Assessing Cognitive Performance Using Physiological and Facial Features: Generalizing Across Contexts

Sensing and machine learning advances have enabled the unobtrusive measurement of physiological responses and facial expressions so as to estimate one's cognitive performance. This often boils down to mapping the states of the cognitive processes underpinning human cognition: physiological responses (e.g., heart rate) and facial expressions (e.g., frowning) often reflect the states of our cognitive processes. However, it remains unclear whether physiological responses and facial expressions used in one particular task (e.g., gaming) can reliably assess cognitive performance in another task (e.g., coding), because complex and diverse tasks often require varying levels and combinations of cognitive processes. In this paper, we measure the cross-task reliability of physiological and facial responses. Specifically, we assess cognitive performance based on physiological responses and facial expressions for 123 participants in 4 independent studies (3 studies for out-of-sampling training and testing, and 1 study for evaluation only): (1) a Pac-Man game, (2) an adaptive-assessment learning task, (3) a code-debugging task, and (4) a gaze-based game. We follow an ensemble learning approach after cross-training and cross-testing with all possible combinations of the 3 first datasets. We save the 4th dataset only for testing purposes, and we showcase how to engineer generalizable features that predict cognitive performance. Our results show that the extracted features do generalize, and can reliably predict cognitive performance across a diverse set of cognitive tasks that require different combinations of problem-solving, decision-making, and learning processes for their completion.

ACM Reference Format:
. 2021. Assessing Cognitive Performance Using Physiological and Facial Features: Generalizing Across Contexts. 1, 1 (January 2021), 41 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Reliably assessing cognitive performance is becoming increasingly relevant in a range of fields encompassing neuroadaptive [80] and critical systems [16, 145], educational technologies [147], operational environments [110], and others. Cognitive performance refers to the overall state of our cognitive functioning, typically comprising of varying levels of cognitive processes, such as attention, memory recall, learning, decision-making, and problem-solving [140]. Over the years, a plethora of cognition measures has been developed for assessing cognitive performance, primarily for the early detection of neurodegenerative diseases, such as Parkinson's, Alzheimer's, and Huntington's. The NIH Toolbox of Cognition Batteries[1] is perhaps the most prominent set of manual cognitive performance measures, incorporating well-established and tested constructs [140]. However, manual cognition measures are cumbersome to employ, require considerable time to complete, and assess one's cognitive capacities on a macro-scale by design [41]. Yet, cognitive performance naturally entails cognitive workload, which is known to influence one's physiological responses[2], such as heart-rate variability (HRV) [53, 132], electro-dermal activity

[1]http://www.healthmeasures.net/explore-measurement-systems/nih-toolbox/intro-to-nih-toolbox/cognition
All hyperlinks last accessed on February 10, 2020.
[2]In this paper, with the term "physiological responses" we refer to skin conductance and photoplethysmography (PPG) sensor data recorded from subjects' wrists, and we acknowledge that the "physiological responses" term is not limited to only such data.

Author's address:

XXXX-XXXX/2021/1-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(EDA) [70], skin temperature [120], but also facial expressions [124]. As a result, automated approaches that utilize physiological responses and facial expressions are gaining popularity in assessing cognitive performance by measuring the produced cognitive workload [9, 30, 52, 66, 90, 111, 117].

1.1 Cognitive Workload: A proxy for assessing cognitive performance

Notably, there is a fine distinction between cognitive performance and cognitive workload that often becomes elusive, particularly when considering that cognitive workload is a natural byproduct of cognitive performance. The Yerkes-Dodson empirical law of arousal is the most prevalent theory describing the relationship between cognitive performance and cognitive workload [149]. The Yerkes-Dodson law, further simplified by Hebb [56], theorizes a non-linear "∩−shaped" curve of increasing (cognitive) performance inline with increasing arousal (workload), leading to an optimal plateau. When cognitive workload is further increased, cognitive performance displays diminishing returns, only to start decreasing rapidly after an empirical threshold is surpassed. The empirical existence of an optimal plateau of productivity is further incorporated in the Flow Theory [26], the experience of mindfulness and complete submersion to the present moment [95]. On one hand, the Flow Theory postulates that when one finds oneself in the "flow zone"—a state of optimal arousal—productivity is maximized. Although the Yerkes-Dodson Law is empirical and the Flow Theory is subjective, they both draw on cognitive workload (arousal) for estimating performance and productivity, respectively. On the other hand, Machine Learning (ML) is tasked with producing affinities and associations even among seemingly unrelated factors, without necessarily unveiling the nature of their relationships. Thus, given the relationship between physiological responses and facial expressions with cognitive workload, and the relationship of cognitive workload with cognitive performance, via ML one can utilize evoked physiological responses and facial expressions to also assess cognitive performance. In other words, we can treat cognitive workload, manifested by physiological responses and facial expressions, as a proxy for estimating cognitive performance.

1.2 Feature Generalizability: Why it matters

Most approaches in literature do not aim at producing generalizable features, and thus they remain inapplicable to other contexts (i.e., context-dependent). As "feature generalizability", we define the extent to which extracted features can predict the same variable—in our case cognitive performance—in different contexts. To this end, feature generalizability is related to "transfer learning" but they differ fundamentally, as we describe later. Prior research has highlighted the importance of generating features that can be generalizable, and particularly in innately-versatile contexts. For example, prior work in music information retrieval considers the generalizability (and simplicity) of features as one of the main criteria for feature selection [112]. More recently, feature generalizability has become relevant when using ML for personality assessment [14]. Feature generalizability also emerges as an important factor when it comes to the automotive context and predicting driver's intentions at intersections [99], as well as students' affect during learning [64]. Likewise, feature generalizability is particularly relevant when dealing with physiological data such as electroencephalography (EEG) for developing Brain-Computer Interfaces (BCIs) [91], and recognizing facial expressions [12]. Now, the field of Ubiquitous Computing naturally involves the introduction of technological interventions to a multitude of contexts. This often implies that certain interventions have to be adjusted to fit a new context. **Knowing a priori which features to compute (and how) for reliably assessing cognitive performance, can save valuable time that would otherwise be allocated to trial-and-error attempts**. Apart from generalizable, the features we engineer in this paper and the methods to compute them, are ideal for hardware with low computational capacities, such as head-mounted displays (HMDs) and smart watches (e.g., VGG16 on a micro-controller [125]).

Here, we contribute to the engineering of generalizable features by expanding the process of modelling cognitive performance to a highly-diverse set of contexts. More specifically, we use 3 datasets of physiological responses and facial expressions that were captured in the context of (1) a Pac-Man game, (2) an adaptive-assessment learning task, and (3) a code-debugging task. Then, we use a 4^{th}, completely new dataset of physiological responses and facial expressions, captured during a gaze-based game, for evaluating the accuracy and generalizability of our features. The cognitive performance of a total of 123 participants was assessed in the form of scores across all 4 study-contexts, where varying levels of problem-solving, memory recall, decision-making, and learning processes manifested. We follow an ensemble learning approach after cross-training and cross-testing with all possible combinations of the 3 first datasets—not by simply merging all datasets—to engineer generalizable features that predict cognitive performance. Finally, we introduce a "**feature generalizability index**" to assess the generalizability of features of physiological responses and facial expressions in a variety of contexts related to cognitive performance. This enables us to decontextualize the knowledge about what works where and how, and contribute to creating strong concepts—constructing knowledge that is more abstracted than particular instances, eventually leading to generalized theories [61] (such as the Flow Theory [26]). In summary, our work makes the following contributions:

- We engineer generalizable features to predict cognitive performance from physiological responses and facial expressions.
- We quantify the generalizability of our features in predicting cognitive performance during problem-solving, learning, and decision-making.
- We propose a "feature generalizability index" (FGI) to quantify the generalizability of features.
- We demonstrate how context-agnostic, cross-training, and cross-testing can yield highly-generalizable features.

2 RELATED WORK

Cognitive performance is not only passively influenced by a plethora of innate factors (e.g., circadian rhythm [137]), but it also affects physiological responses and facial expressions as a result of exhibiting cognitive workload [70, 120, 124, 132]. Thus, assessing and eventually improving cognitive performance, with the use of physiological data, has been the focal point of numerous studies in the intersection of Ubiquitous Computing, Human-Computer Interaction (HCI), Educational Technologies, and Neuroergonomics fields. Next, we report on prior research that utilizes physiological responses and facial expressions for assessing and improving specific cognitive processes or cognitive performance overall.

2.1 Physiological Responses and Cognitive Performance

A large body of research is dedicated to monitoring cognitive workload, engagement, or enjoyment, drawing among others on Flow Theory [26]. For example, Rissler et al., build on Flow Theory for developing so-called "flow-classifiers" that use cardiac features for classifying flow states during an invoice matching task [105]. Schaule et al., utilized consumer smartwatches for measuring office workers' physiological responses for inferring cognitive workload and deciding when the time is right to be interrupted [114]. Their approach involved a feature vector generated among others from time and frequency features of HRV, EDA, and skin temperature. In the same guise, Goyal and Fussel employed the Q Sensor by Affectiva[3] for monitoring EDA during collaborative tasks and managing interruptions [48]. In particular, they calculated the direction of intensity of change in the average EDA phasic amplitude as a feature to decide over one's interruptibility.

Gjoreski et al. also used low-cost wrist-worn devices for monitoring cognitive workload based features extracted from physiological responses such as HRV and EDA, in conjunction with the established self-assessment NASA-TLX method [47]. Similarly, Kosch et al., used EDA recorded from the Empatica E4 wrist-worn device for monitoring cognitive workload during a manual assembly tasks with 2 different assistive systems [72]. Using

[3]https://www.affectiva.com/

Bayesian Repeated Measures ANOVA and the NASA-TLX method, they concluded that EDA is an objective measure for workload monitoring during assembly tasks. Mirjafari et al., was able to collect among others physiological responses from over 550 office workers in the period of 2–8 months for assessing performance in the workplace [83]. The authors found significant correlations between high performance in the workplace and regular heart-beat rates during the weekdays. In a study that involved over 100 drivers, Solovey et al., showcased that features from physiological responses improve the detection of increased cognitive workload when driving with an accuracy of 90 % [122]. More recently, psycho-physiological sensing for assessing cognitive workload and operational performance has also been proposed for monitoring the cognitive states of an aerospace crew [146]. Typically, HRV monitoring with wrist-worn devices is performed via photoplethysmographic (PPG) sensors embedded in the back of the devices touching the skin. In a different approach, Zhang et al., employed a PPG-based method to measure cognitive workload (mental stress) during touch interactions with an infrared touchscreen [154]. By utilizing HRV features measured with PPG, they were able to classify cognitive workload with an accuracy of 97 % and 87 % during static and interaction testing, respectively.

Physiological responses have also been extensively utilized for measuring engagement and enjoyment in gaming experiences. For example, EngageMon is a multi-modal engaging sensing system that combines a wide range of physiological and contextual data for assessing engagement during mobile gaming [65]. Among others, the authors utilized features extracted from the HRV and EDA physiological responses, combined with features from video and mobile usage to achieve an average accuracy of 87 % in estimating engagement. Tognetti et al., utilized physiological responses such as Electrocardiography (ECG) data, EDA, Blood Volume Pulse (BVP), and respiration data captured with the ProComp Infiniti[4] device during a racing game for gauging enjoyment [135]. In an alternative approach, Tan et al., utilized the think-aloud method in conjunction with Electromyography (EMG) data collected with the ProComp Infiniti for understanding video-game experiences [130]. The authors did not apply any ML technique, but instead classified manually the EMG peak data in 4 different categories concluding that physiological data can be used as “anchors” in labelling think-aloud reports.

Learning is also disrupted by approaches that utilize physiological responses for gauging engagement, monitoring learning performance, and adapting learning difficulty. Di Lascio et al., used the Empatica E4 physiological-monitoring wristband for assessing the engagement of students during lectures [29]. Except for monitoring arousal, the authors used EDA data for designing features that characterize the “physiological synchcrony” between the students and the teacher for better estimating engagement in the classroom. In a followup work, Gashi et al., investigated the notion of “physiological synchcrony” predicted by EDA features for estimating engagement between presenters and the audience in conjunction with subjective self-reporting measures [42]. Ghiani et al., used EEG and eye-tracking data for creating attention rules based on which they throttle information presentation for facilitating learning [45]. Tamura et al., utilize simple EEG amplitude features of the beta band in combination with eye-tracking and subjective assessments to detect difficult to comprehend content during e-learning [129]. Radeta et al., employed the Empatica E4 for acquiring EDA measurements to compare between 2 interactive learning experiences for kids [103]: a mobile game vs. animated storytelling. The authors were able to quantify and link learning for both experiences to EDA peaks.

2.2 Facial Expressions, Emotions, and Cognitive Performance

Emotions influence arousal and affect, bearing important effects on productivity and cognitive performance [26, 95], and can be reflected in physiological responses [51]. Nevertheless, facial expressions are perhaps the most reliable indicator of emotion, as Ekman has shown [37]. Thus, facial expressions have been used either in isolation or in conjunction with physiological responses for assessing mood and cognitive performance. Babu et al., propose a multi-modal approach for measuring task-based cognitive performance that utilizes both facial

[4]http://thoughttechnology.com/index.php/procomp-infiniti-333.html

expressions and EEG data [9]. They used the VGG-19 network to generate a set of feature maps from images of participants performing a sequence learning task, and a Convolutional Neural Network (CNN) for predicting emotions. By also incorporating EEG input, they were able to assess task-based cognitive performance with an accuracy of 87.5 %. In the first “audio-visual+” emotion recognition challenge, Ringeval et al., merged video of facial expressions with physiological data for detecting affective dimensions of arousal [104]. The authors describe how they extracted features from videos of facial expressions using the Supervised Descent Method (SDM) [148], and features from physiological responses (EDA and HRV), computing among others the spectral entropy and the first order derivative.

During learning, emotions play a very integral role. Happiness is related to high prospective success, whereas anger is related to retrospective failure, and sadness to high negative activity [97]. On one hand, exhibiting happiness/joy results in novel and creative actions [40], while positive emotions also promote the engagement in meta-cognitive processing, beneficial for long term learning [77]. On the other hand, negative emotions result in focusing on environmental-specific details [15], and may reduce elaboration [96]. Moreover, negative affect has been associated with lower learning goals [82], whereas positive affect with the interest in a given topic [3]. Thus, bearing in mind the innate connection between emotions and facial expressions, a sizeable body of research is dedicated to assessing learning performance through emotions inferred from facial expressions [10, 36, 49, 50]. In multiple instances, D’Mello et al., collected facial expressions of students, while interacting with the “AutoTutor” learning system [25], and played their facial expressions back to them asking them to annotate their emotions during their prior interactions with the learning system [33–35]. In this way, the authors were able to model the transition likelihood among the affective states of boredom, flow (engagement), and confusion during learning. Baker et al., were perhaps the first to adapt an automated approach for detecting affective states during learning by using a large dataset with manually-labelled affective states of students that also contained their facial expressions [27]. The authors used eight common classification algorithms (e.g., J48, decision trees, Naive Bayes, etc.) but with mixed results. Similarly, Whitehill et al., assembled a dataset comprised of videos from facial expressions of 34 undergraduate students, interacting with a software that trains their cognitive skills, along with their performance scores [142]. The dataset was then manually labelled by researchers producing 4 levels of engagement. The authors then applied binary classification techniques to automatically classify engaged from non-engaged students from their facial expressions, using Boost(BF), Support Vector Machine (SVM), and Multinomial Logistic Regression (MLR), with the manual engagement values and the facial expressions coded in Action Units (AUs). Notably, the authors considered the generalization issue of facial classifiers when it comes to classifying facial expressions of people with dark skin colour. To rectify this, they opted for diversifying their dataset by including African-American, Asian-American, and Caucasian-American participants, and cross-testing between different populations. Their results showed that Boost(BF) classifier generalized well to subjects within the same population but not to subjects of a different population [142].

Recently, commercial and open-source software approaches have emerged for facilitating the automatic emotion assessment from facial expressions. For example, FaceReader is a commercial automated facial-coding software that displays good accuracy when compared with human emotion recognition from facial expressions [76]. OpenFace is an open-source facial behaviour analysis toolkit that implements facial landmark detection and tracking, as well as eye-gaze and head-pose estimation [11]. In particular, AU recognition has been tested in multiple publicly available datasets, displaying better performance in videos of facial expressions than in pictures. Either experimental, commercial, or open-source, approaches that infer emotions from facial expressions for assessing aspects of cognitive performance are seldom tasked with producing generalizable features [142]. The same trend is observed when having a look at approaches that utilize physiological responses for the purpose of assessing cognitive performance. However, producing generalizable features for reliably assessing cognitive performance in diverse contexts paves the way for designing the cognition-aware systems of the future [17, 30].

2.3 Feature Generalizability vs. Transfer Learning

"Transfer Learning" (TL) is a machine learning concept related to our work, but fundamentally different to feature generalizability. TL uses the produced model from one task to improve performance at a rapid pace for another related task [92]. There are two main methods for accomplishing this [93, 141]: (a) develop a new model, and (b) use a pre-trained model. The first method involves the selection of a related prediction problem with a large set of training data available. The new model is developed on related training data, and then the entire model, or part(s) of it, is (are) used in the original prediction problem [151]. The second method assumes a pre-trained model, and reuses or adjusts its original parameters to fit the targeted prediction problem [92]. Essentially, transfer learning is about finding the feature set that will work both for the related and target contexts [151]. However, feature generalizability is not the main aim of TL. In fact, TL solely focuses on optimizing the prediction outcome in the target context. That is, the model trained in the related context (features and their relation to the predicted variable) is reused as is (a), or the model can be tuned to fit the target context (b). Conversely, feature generalizability does not necessarily optimize the prediction outcome for any of the selected contexts. Additionally, the outcome is a set of features that are empirically deemed to be useful across all the selected diverse contexts. Finally, although TL requires considerably large datasets for training the base model, large datasets is not a requirement for achieving feature generalizability.

3 STUDY DESIGN

Our aim is to engineer generalizable features that predict cognitive performance from physiological responses and facial expressions by drawing on 4 independent study datasets: (1) a Pac-Man game, (2) an adaptive-assessment learning task, (3) a code-debugging task, and (4) a gaze-based game only for evaluation purposes. In all studies, intrinsic facets of cognitive performance were central to the completion of the task at hand, and were objectively assessed by performance indices (scores). In the studies involving games (i.e., 1 and 4) the score is related to skill-acquisition and in the educational studies (i.e., 2 and 3) the score is related to problem-solving capacities. In particular, we theorize that the 1^{st} study (Pac-Man game) involves problem-solving, decision-making, and learning. The 2^{nd} study (adaptive-assessment of learning) involves problem-solving, decision-making, and memory recall that trigger learning. The 3^{rd} study (code debugging) entails a combination of problem-solving and learning. Finally, the 4^{th} study also involved problem-solving, decision-making, and learning, and its sole purpose was evaluating our engineered features in a completely new context. During all 4 studies, physiological responses and facial expressions were collected, along with the corresponding performance index (score) for each participant. For all 4 studies, we have obtained the appropriate ethical approval (details hidden for anonymization). In all 4 studies, the data (facial and physiological) was collected using Empatica E4 wristband and a Logitech web camera. Moreover, in the 4^{th} study (i.e., evaluation of the generalized features), participants interacted with the game via their eye gaze, without touching the input devices and/or the screen.

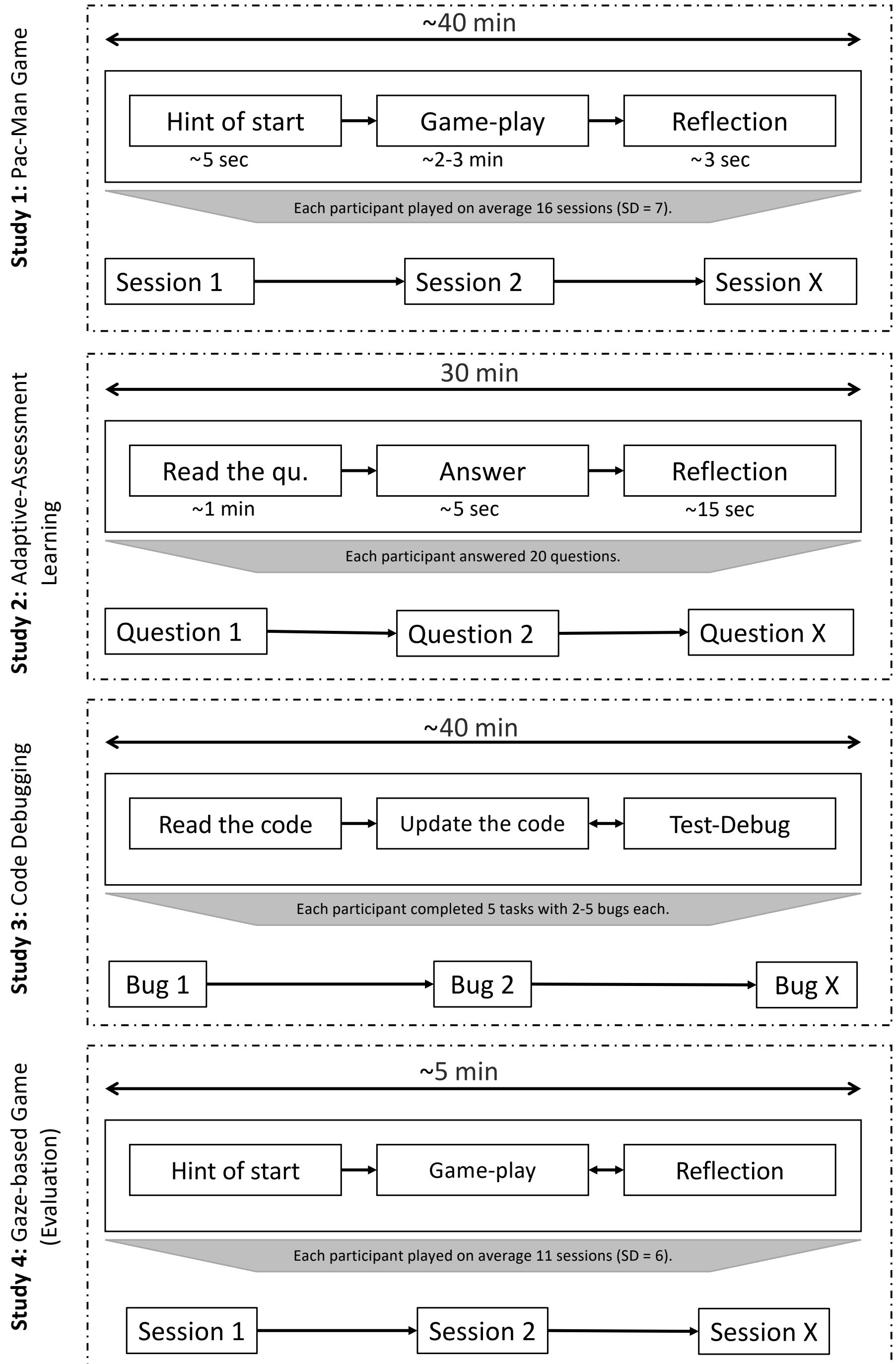

Fig. 1. Protocol of the 4 studies, including the gaze-based game (Study 4) for evaluating our features in a completely new context.

3.1 Study 1: Pac-Man Game

This study is a controlled experiment focusing on skill acquisition. Skill acquisition (or movement-motor learning [39]) loosely encompasses motor adaptation, problem-solving [138], and decision-making [73, 144]. Skill acquisition consists of the memorisation of an internal representation of a movement (conceptualised as a motor schema) [133]. Thus, skill acquisition also involves learning. When one receives guidance verbally or one rehearses mentally the skill to be acquired, one exhibits cognitive workload, indicating the manifestation of higher cognitive processes [133]. To maintain a simple learning curve, we developed a Pac-Man, a time-tested game that has been used to test motor skills in the past [87]. Pac-Man was developed by applying all the typical game-play elements (e.g., enemy sprites and the maze—see Fig. 2), while providing 3 lives for each session. The game was controlled by the 4 arrow buttons of the keyboard, and was developed to log every keystroke performed by the user. The difficulty of the game increased from one session to another by increasing the sprite-movement speed.

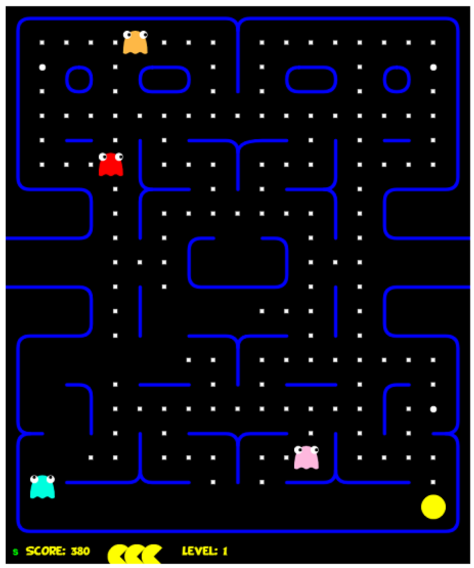

Fig. 2. Study 1: the custom-made Pac-Man game. The basic design principles of the game is minimalist design and a highly-immersive game environment.

3.1.1 Participants. We recruited a total of 17 healthy participants (7 females) aged 17–49 years ($M = 32.05, SD = 8.84$) over May 2018. The participants were recruited from the participant pool of a major European university. All participants were familiar with the game, but none of them had played the game in the previous 2 years. Prior to completing the trials, the participants were informed about the purpose and the procedure of the experiment, and of the harmlessness of the equipment involved. We compensated the participants with a movie ticket upon the completion of the study.

3.1.2 Protocol. The experimental design of the Pac-Man study was a single-group time series design [107] with continuous (repeated) measurement of a group exposed to the same experimental intervention. Each participant played on average 16 game-sessions ($SD = 7$), until their allocated time ran out. Each game-session started with 3 lives and ended when the participant lost all 3 lives. For each level in a game-session, the speed of the ghosts-sprites increased. Fig. 1 showcases the protocol of our experiment. Each participant was given a 5-seconds break before starting the next session. Each session was completed in 2–3 minutes, after which the participants had a 2–3 seconds of reflection period, looking at their game score.

3.1.3 Procedure. Upon obtaining consent, the researcher escorted the participant to the User Experience (UX) room with a comfortable chair facing a large computer monitor. The participant wore the Empatica E4 wristband, while the researcher connected and calibrated all the data collection devices (i.e., E4 wristband and camera). The wristband data streams were calibrated using the built-in calibration procedure available in the Empatica mobile application. The researcher explained the mechanisms of the game and the respective keyboard functions, double-checked the data collection devices, and exited the room. The participant had ~40 minutes to master the game and achieve a score that was as high as possible.

3.1.4 Performance. At the end of each game-session the participants received a score that was considered as their performance in that session. Thus, we use the game-score as an indicator of cognitive performance.

3.2 Study 2: Adaptive-Assessment Learning

The 2nd study also took place in controlled settings and focused primarily on learning, by also encompassing the cognitive processes of problem-solving, memory recall, and decision-making. Students' responses and system usage logs were collected with LAERS [94], a web-based implementation of a layered architecture for testing systems. The version of LAERS employed in this study consists of (a) an assessment interface, (b) an adaptation mechanism, (c) a tracker that logs the students' usage data when interacting with the system, and (d) a database storing information about students' performance and the test-items.

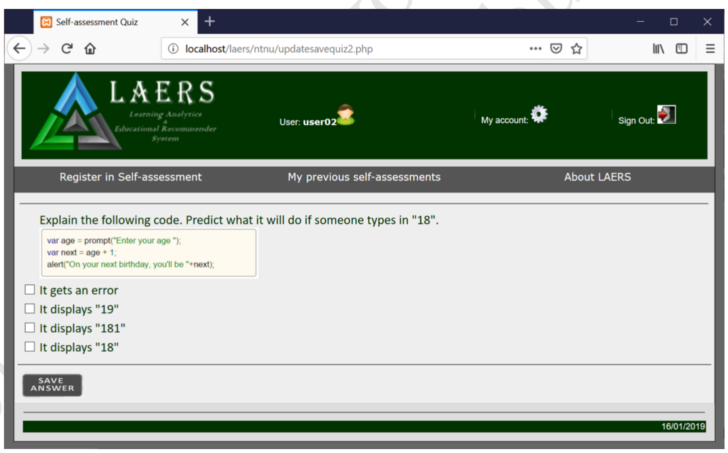

Fig. 3. Study 2: The LAERS self-assessment interface featuring a test-item that requires solving a short coding exercise so that it can be answered.

The assessment interface displays the test-items, in the form of multiple choice questions, which are delivered to students one by one (see Fig. 3). The adaptation mechanism selects the next most appropriate test-item to deliver to the student, according to the correctness of the student's response to the previous test-item, and the discrimination capacity of the test-items, by drawing on the Measurement Decision Theory (MDT) [109]. The tracker logs the students' response time, dividing it to time on correctly- and time on wrongly-answered test-items. Finally, the system also calculates and updates the test score according to the correctness (0/1) of the student's answer for each test-item.

3.2.1 Participants. The study was conducted at a controlled computer lab, equipped and furnished for the needs of the experimental process, over October 2018. We recruited a total of 32 undergraduate students (15 females) aged 18–21 years ($M = 19.24, SD = 0.831$) from the pool of a European University. All participants were enrolled in an online adaptive self-assessment procedure for the Web Technologies course (related to front-end development). The participants undertook the self-assessment task individually for a period of ~30 minutes each.

3.2.2 Protocol. The experimental design of the adaptive assessment study was a single-group time series design [107] with continuous (repeated) measurement of a group exposed to the same experimental intervention. Each participant answered 20 questions, in about 30 minutes. Each test-item provided 2–4 possible answers but only one of them was correct. Some test-items required factual and/or conceptual knowledge to be answered, whereas others were puzzles (i.e., short coding exercises), thus requiring procedural knowledge to be solved [74]. Each session lasted from the display of test-item until providing an answer (~1 min). Fig. 1 presents the protocol of this experiment. Each participant was shown the correct answer before moving to the next test-item. In the end, a list containing the test-items and their answers was shown to the participants, and they had 2–3 minutes to reflect on their performance.

3.2.3 Procedure. Prior to the experiment, all participants signed an informed consent form that detailed the procedure, authorising the researchers to use the data collected for research purposes. After granting their consent, the participants had to wear the E4 wristband, and all data collection devices (i.e., wristband and camera) were tested. Furthermore, the participants had to answer to a pre-test questionnaire that assessed their goal-expectancy from the upcoming self-assessment. Next, the actual adaptive self-assessment experiment commenced, with the students providing their answers to the test-items. In the end of the procedure, the test score was made available to the participants, along with their full-test results, including all the test-items to which they had responded, their responses, the correctness of their responses, and the option to check the correct answers to the test-items that they had submitted wrong answers. This was intended for self-reflection purposes. Finally, the participants were compensated with a movie ticket upon the completion of the study.

3.2.4 Performance. Each response to a test-item in an individual session was given a correctness label (0/1). This was considered as the performance measure for this experiment.

3.3 Study 3: Code Debugging

Drawing on Katz and Anderson's conceptualization of a debugging process [68], we decided to engage the debugging process as a case of troubleshooting featured in 4 steps: (1) understand the problem, (2) find the bug, (3) fix the bug, and (4) test the code. In this study, we postulate the manifestation of cognitive processes that involve problem-solving and learning. In fact, debugging is more related to procedural knowledge than it is to factual or conceptual knowledge [74]. We designed and implemented a debugging task to collect a fine-grained multi-modal dataset and explore the features associated with cognitive performance in the debugging process. The main task assigned to the participants was debugging a Java class named "Person" (that implements "parent-child" relationships), accompanied with five debugging tasks (i.e., questions), written right after the code, and presented as a part of the main method.

3.3.1 Participants. The study was conducted in the controlled settings of a computer lab at a European university with 46 students (8 females) over the Spring semester 2019. Participants were recruited from all study years of the computer science major of our University via an e-mailing list. We specifically did not recruit participants in their 1st year, since they had not taken an object-oriented programming (OOP) course yet. All participants had used Eclipse Integrated Development Environment (IDE) during their OOP course. For their participation in the study, participants received a gift voucher equivalent to $35.

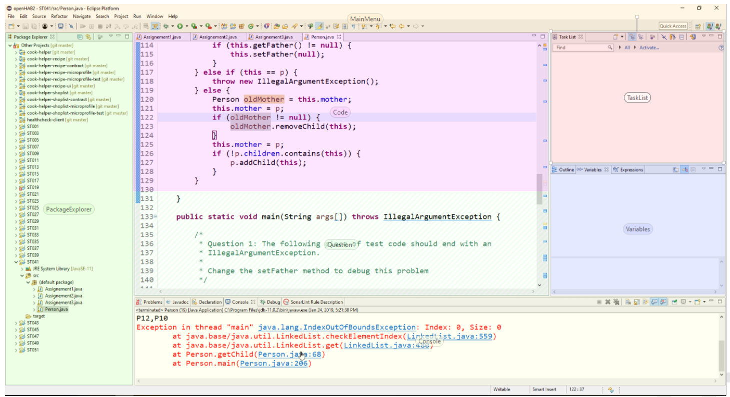

Fig. 4. Study 3: the Eclipse IDE and the panels available to the participants.

3.3.2 Protocol. Similar to the previous studies, the research design of the code-debugging study is a single-group time series design [107] with continuous (repeated) measurement of a group exposed to the same experimental intervention. Each participant was requested to complete 5 debugging tasks with a total duration of ~40 minutes. Each task was composite, requiring the debugging of 2–5 "bugs" in order to be completed. Fig. 1 displays the protocol of this experiment. The participants were allowed to modify the code as many times as they desired. In the end, one of the researchers explained the participants which were the remaining bugs and how to fix them.

3.3.3 Procedure. Upon arrival in the laboratory, the participants signed an informed consent form. Next, the lead researcher placed the E4 wristband on their wrist, and all data collection devices (i.e., wristband and camera) were tested. The wristband data streams were calibrated using the built-in calibration procedure available in the Empatica mobile application. Before the actual study commenced, the participants were asked to complete 3 small debugging assignments (easy, medium, and difficult) within 20 minutes. This pre-test was intended for assessing the debugging expertise of the participants. Then, the participants were given 40 minutes to complete the 5 debugging tasks (i.e., questions) presented as part of the main method in the "Person" class. The provided code assumed, but failed to ensure, consistent object relationships (e.g., "a mother of a child is female"). The 5 debugging tasks were incremental. Thus, the participants could not start working on the second task if they had not successfully completed the first one. The code for the main debugging task contained no syntax errors, and the participants were informed about this fact.

3.3.4 Performance. At the end of the experiment, the participants were assigned 5 individual scores based on the number of bugs they fixed in each debugging task. This was the performance measure for this experiment.

3.4 Study 4 (feature evaluation only): Gaze-based game

Similar to the Pac-Man game, this study is also a controlled experiment focusing on skill acquisition, including problem-solving [138], and decision-making [73, 144]. However, all interactions are explicitly performed through eye-gaze, and thus we assume an extent of ocular motor adaptation as part of skill acquisition [116]. **Most importantly, we theorize that the context of this study more closely aligns with previous studies and applications in the field of Ubiquitous Computing, involving pervasive displays and gaze-based interaction** (e.g., [123]). The gaze-based game is called Xtreme Yoga, and it is a shooting game we developed for the stationary Tobii eye-tracker. In the game, a player controls an avatar with 3 lives that avoids randomly appearing "knights" and the projectiles they launch. The avatar can move in all directions, and its movement entirely relies on the player's eye-gaze. An always-visible white circle indicates where the player's eye-gaze is focused (Fig. 5a). Additionally, a player can focus on the avatar to activate a defensive shield (Fig. 5d and 5f), or focus on a “knight” at whom to launch a projectile (Fig. 5c). The game ends when a player loses all 3 lives. **The dataset of this study was explicitly used only for evaluating the generalizability of the features we engineered based on the previous studies described**.

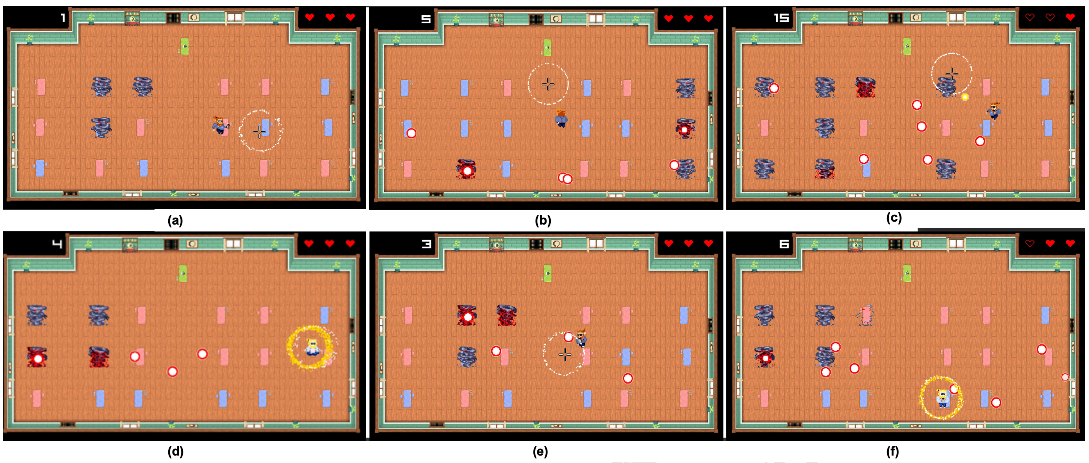

Fig. 5. Study 4: the different stages in the game.

3.4.1 Participants. We recruited 28 healthy participants (8 females) aged 8–14 years ($M = 10.00, SD = 1.38$) over November 2019. The participants were recruited from a classroom of a major public school in a European city. None of the participants were familiar with the game or its gaze-based controls. Prior to completing the trials, the participants were informed about the purpose and the procedure of the experiment, and of the harmlessness of the equipment involved. We compensated the participants with a gift coupon equivalent to $11 upon the completion of the study.

3.4.2 Protocol. Similar to Studies 1, 2, and 3, the experimental design of the gaze-based game study is a single-group time series design [107] with continuous (repeated) measurement of a group exposed to the same experimental intervention. Each participant was requested to play multiple sessions of the game, with a session duration of ∼5 minutes. In each game session, participants used their eye-gaze to avoid projectiles, raise shields, and attack an enemy to increase their overall score. Fig. 1 displays the protocol of this experiment. The participants were allowed to play the game as many times as they desired. On average, each participant completed 11 game sessions ($SD = 6$).

3.4.3 Procedure. Prior to their arrival in the laboratory, the participants’ parents signed a parental / guardian consent form at home. Next, the lead researchers placed the E4 wristband on the wrists of the participants. The wristband data streams were calibrated using the built-in calibration procedure available in the Empatica mobile application. Participants’ facial expressions were recorded with a webcam. Before the actual study commenced, the participants were asked to play one training round so that they familiarize themselves with the gaze-based controls and the game setup. The researcher explained the mechanisms of the game and the respective gaze-controlled functions, double-checked the data collection devices, and exited the room. Then, the participants were asked to play as many games as they desired. Each game session had 3 player lives, once all were lost the participants could restart the game.

3.4.4 Performance. At the end of each game session the participants received a score for their performance in that game session. The score increased the longer a player kept the avatar alive, and the more enemies a player terminated. Thus, we use the game-score as an indicator of cognitive performance. The score was set back to 0 each time a new game session started.

4 ANALYSIS

To engineer generalizable features from physiological responses and facial expressions, we utilise the datasets collected in Studies 1, 2, and 3, and validate the features using the data from Study 4. The total sample size for all studies was 123 participants. To identify the generaliszable features, first we apply standard data pre-processing techniques: denoising, filtering, smoothing. Next, we perform a common feature engineering process to extract the features from the raw signals, and we then reduce the feature space either by applying a feature selection technique (keeps the selected features in their original form), or by using a dimensionality reduction algorithm (creates new dimensions using certain combinations of the original features). The final step is to apply an ensemble of prediction algorithms to predict cognitive performance. Fig. 6 summarises the overall process applied in our analysis.

To identify generalizable features, we conduct an exhaustive search of possible analyses and data combinations. Therefore, in the remainder of the paper we use the term “**pipeline**” to refer to a unique combination of: studies (i.e., Pac-Man, Adaptive-Assessment Learning, and Code Debugging) and data (i.e., physiological and facial expressions) as input, extracted features (e.g., deep features, action units, FFT, LPC, etc.), either feature selection (e.g., LASSO) or dimensionality reduction (e.g., Kernel PCA), and ensemble prediction models (e.g., Support Vector Machines, Gaussian process models, etc.). We opt to test both *feature selection* and *dimensionality reduction* methods, since in the attempt to engineer generalizable features, there is no empirical / theoretical grounding for any of the 2 methods to perform better. Notably, each pipeline uses either the feature selection or the dimensionality reduction, never both. **In a nutshell, a “pipeline” is a unique combination of data inputs, selected features or reduced feature sets, and prediction models**.

A total of 156 pipelines was assembled and tested in our analysis. Each pipeline receives one of the three data types as input: (1) physiological data, (2) facial data, or (3) both (see Section 4.1). The data from the E4 wristband and the facial videos are first pre-processed to remove the noise and bias from known sources, including hand movement and camera white-balancing (see Section 4.2). The features are extracted based on the data type used in each pipeline: signal processing features from physiological data—action units and deep features from facial data (see Section 4.3). Once the features are extracted, they serve as input to either the feature selection (LASSO, linear or RF, non-linear, see Section 4.4), or the dimensionality reduction (PCA, linear or kernel PCA, see Section 7.6 of the Appendix). Features selected via either branch comprise yet another pipeline. Next, the selected features (in the case of feature selection), or the modified space (in the case of dimensionality reduction), serve as input to the ensemble learning setup with seven predictors (SVM—linear, radial, polynomial; model tree M5; GPM—linear, radial, polynomial, see Section 4.6). The weighted average, after performing 10-fold cross-validation and out-of-sample testing, yields the final prediction over cognitive performance drawing on data from all 3 independent studies. For engineering generalizable features, we also perform “out-of-study testing” (i.e., leave-one-task-out), testing the engineered features on entirely different datasets from the ones on which they were trained (see Section 4.7). We also introduce a feature generalizability measure, based on which we compare our pipelines (see Section 4.8), and we benchmark the generalizability of the top performing features in a completely novel context (Study 4: Gaze-based game—see Section 4.9).

Finally, we point out that for the 4 studies, and the 4 respective tasks presented in this paper, cognitive performance is calculated slightly differently. For the games, such as Pac-Man and the gaze-based game, there is no theoretical upper limit for the score. Conversely, the scores are upper-bounded in the adaptive assessment and the code debugging tasks (i.e., having all the tasks correct and achieving the maximum score). Yet, even though the performance measurements have different ranges, there is a key similarity across all the tasks: high performance requires a certain level of (i) skill, (ii) attentional processing, and (iii) cognitive processing across all tasks. In addition, Table 1 presents the mean values of the cognitive performance, their standard deviation, and the results coming from a chi-square comparison on their distribution. Table 1 indicates that the cognitive-performance slightly varies across the different tasks, but with no statistically significant difference. Moreover, the mean values and their standard deviations depict that there was a healthy distribution of the cognitive performance in each of the tasks (i.e., we did not have a very difficult or very easy task). Another commonality between the 4 tasks is that for the user to attain high cognitive-performance score, they need to devote the required levels of attentional and cognitive processing. **This paper is an effort to identify those facial and physiological features that generalise across different contexts to encode these attentional and cognitive processing levels that are associated with task-based cognitive performance.**

Table 1. The second column depicts the mean and standard deviations for the cognitive-performance measures (normalized using MinMax) from the 4 studies. The third-sixth columns depict the results of chi-square tests for the distributions of the cognitive performance measurements of the 4 studies. The number indicates the chi-square statistic, and the number in the parentheses the corresponding p-value.

	Mean (SD)	Pac-man	Adaptive Assessment	Code Debugging	Gaze-based Game
Pac-man	0.35 (0.29)	–	18.61 (0.54)	21.25 (0.38)	25 (0.20)
Adaptive Assessment	0.48 (0.26)	–	–	25.83 (0.41)	23.61 (0.54)
Code Debugging	0.59 (0.36)	–	–	–	28.75 (0.27)
Gaze-based Game.	0.32 (0.28)	–	–	–	–

4.1 Cross-Study Data Collection Setup

We collected sensor data from 2 different sources: (a) the Empatica E4 wristband, and (b) a video camera.

- **E4 wristband: To record physiological data we use the Empatica E4 wristband.** Participants wore the wristband on the non-dominant hand. Four different measurements were captured: (1) heart rate variability (HRV) at 1 Hz, (2) electrodermal activity (EDA) at 64 Hz, (3) body temperature at 4 Hz, and (4) blood volume pulse (BVP) at 4 Hz.
- **Video camera:** Given the fact that we expected participants to exhibit minimal body and gesture activity during all the 4 studies, the video recording focused on their face. We use a Logitech Web cam capturing video at 30 FPS. The webcam focus was zoomed 150 % onto the faces of the participants. The video resolution was 640×480 pixels.

4.2 Data Pre-processing

We pre-processed the following types of data as follows:

- **Physiological data:** A simple smoothing function was used to remove any unwanted spikes in the time series in the 4 data streams originating from the E4 wristband (HRV, EDA, Skin Temperature, and BVP). This was a simple running average with a moving window of 100 samples, and an overlap of 50 samples between two consecutive windows. Physiological data, such as HRV, BVP and skin temperature, are susceptible to many subjective and contextual biases. These biases include: time of the day, physical health condition, gender, age, overnight sleep, and others. All 4 data streams were normalised using the first 30 seconds of the data to remove the subjective and contextual biases from the data.

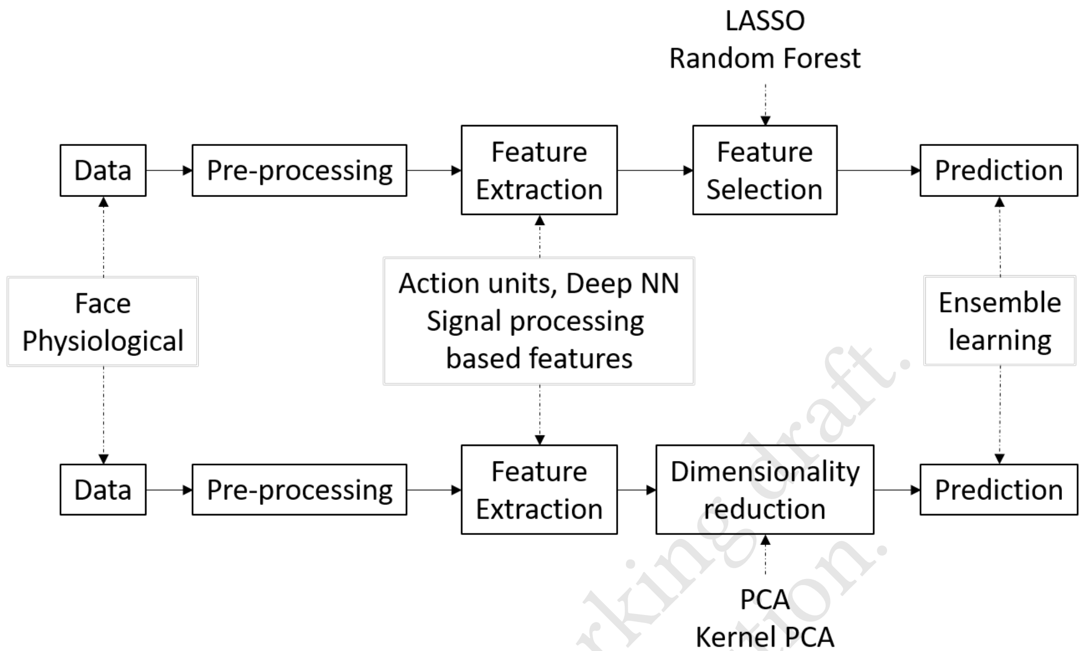

Fig. 6. General pipelines (top: with feature selection; bottom: with dimensionality reduction) for the prediction of cognitive performance.

- **Facial data:** For most of the frames in the video recordings, only one face was visible. However, sometimes the lead researcher appeared in the field of view of the camera. Due to the settings of the experimental space, the researcher could only appear to the right side of the participant. Moreover, the algorithm in the OpenFace face recognition library [6] assigned each face in the frame a unique identifier from left to right. This means that in the frames, where both the researcher and the participant were present, the participant's face unique identifier was always zero. For frames with 2 faces (as this was the highest number of faces in any frame), the researcher's face that had a unique identifier value of 1 was systematically removed.

4.3 Feature Extraction

4.3.1 Features from physiological data. We computed the following features from the physiological data streams (EDA, HRV, skin temperature, and BVP). These features are extracted based on the recent approaches for feature extraction using both the time [48, 72, 118, 146] and the frequency [44, 119, 153] domain properties of the data.

- **Value histogram:** We computed the mean, standard deviation (SD), skewness, kurtosis, and median of the value histogram of the 4 data streams.
- **Power spectral histogram:** The power spectrum of a time series describes the distribution of power into frequency components composing that signal. Once the frequency components are computed, they can be represented as a histogram (Power Spectral Histogram). We computed the mean, SD, skewness, kurtosis, and median of the Power Spectral Histogram. Fig. 12 displays the individual differences among features extracted with the power spectral histogram process.

- **ARMA:** An ARMA process combines the auto-regressive and the moving average features. More precisely, $X(t)_{t\in\mathbb{Z}}$ follows an ARMA process if for every t the random variable X_t satisfies

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{1}$$

  In order for these equations to define a covariance stationary causal process (a process that depends only on the past innovations), the coefficients must be $\left|\phi_j\right| < 1$ and $|\theta_i| < 1$. Moreover, ϵ models the residual noise. Fig. 10 displays the individual differences among features extracted with the ARMA process.
- **GARCH:** GARCH models are similar to AutoRegressive Moving Average (ARMA) models but they are applied to the variance of the data instead of being applied to the mean [4, 38, 69, 78, 113]. GARCH processes $X(t)_{t\in\mathbb{Z}}$ take the general form

$$X_t = \sigma_t Z_t, t \in \mathbb{Z} \tag{2}$$

  Where σ_t, the conditional deviance (so-called volatility in finance), is a function of the history up to time $t-1$ represented by H_{t-1} and $(Z_t)_{t\in\mathbb{Z}}$ a strict white noise process with mean zero and variance one. We assume that Z_t is independent of H_{t-1}. Mathematically, σ_t is H_{t-1} measurable, where H_{t-1} is a filtration generated by $(X_s)_{s\le t-1}$, and therefore

$$X_t|H_{t-1} = \sigma_t^2 \tag{3}$$

  The series (X_t) follows a $GARCH(p,q)$ process if for all t

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{k=1}^{q} \eta_k \sigma_{t-j}^2, \alpha_j, \eta_k > 0 \tag{4}$$

  The condition on the parameters, $\alpha_j = 1 \ldots p$ and, $\eta_k = 1 \ldots q$ for the GARCH equations to define a covariance stationary process with finite variance is that

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \eta_k < 1 \tag{5}$$

  The rationale behind equation 4 is that, first, opposite to AutoRegressive Moving Average (ARMA) models, which are models for the conditional mean, the GARCH is a model for the conditional standard deviation. By "conditional" we mean "given the history up to time t", that is given H_{t-1}. Second, the model shows that more persistence is built into the variability. In other words, GARCH models the variance at time t in the time-series as the linear combination of the history of variances up to time $t-1$. For more details see [131]. The coefficients $\alpha_0 \ldots \alpha_p$ and $\eta_1 \ldots \eta_p$ can be estimated by maximizing a likelihood function. The most popular GARCH model is $GARCH(1,1)$, that is, $p = q = 1$ in (3) meaning that the current action variability is explained by the latest action and the latest action number only (lag time of one). Fig. 11 displays the individual differences among features extracted with the GARCH process.
- **Linear Predictive Coding (LPC):** This is a way of coding the spectral envelope of the signal. LPC is mostly used to perform lossless compression of the signals [71, 85, 152], however it has recently been used to analyse the quality of the signal as well [128]. LPC estimates the amplitude for signal x_n as:

$$\hat{x}_n = -\alpha_1 x_{n-1} - \alpha_2 x_{n-2} - \alpha_3 x_{n-3} \ldots - \alpha_p x_{n-p} \tag{6}$$

- **Linear Spectral Frequency Coding (LFSC):** LPC is susceptible to high peaks in the signal [7], hence we also compute the LSFC for the physiological data that improves upon this shortcoming of the LPC [67]. Fig. 13 displays the individual differences among features extracted with the LPC and LFSC processes.

4.3.2 Features from facial data. The most common feature extraction techniques used in the literature are Action Units (AU) [28, 50, 106], and deep features [9, 83, 126]. Thus, for ensuring we have extracted all the potential features, we applied both techniques in our feature extraction stage.

- **Action Units (AU):** Using facial data (videos of facial expressions), we elicited expressions and produced features from different face regions (eyes, nose, mouth, jawline). Following best practices in literature, we extracted the facial Action Units (AUs,[24]) using the OpenFace library [6]. Fig. 7 shows the AUs detected in this study. We detected these AUs for each frame in the video. OpenFace provides a floating point value between 0 (nothing detected at all) and 5, based on the intensity of each AU detected. Fig. 9 displays the individual differences among extracted AUs.

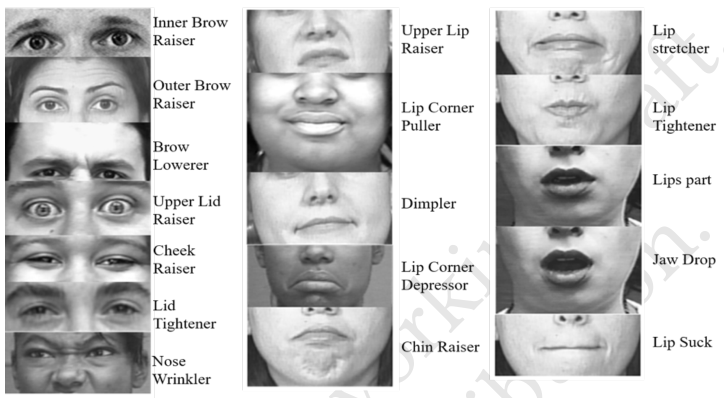

Fig. 7. Action Units (AU) correspond to the fundamental actions of different facial muscles or group of facial muscles [11].

- **Deep features**[5]**:** Using the deep neural network architecture by Simonyan and Zisserman [121], we extracted the "deep features" in the following steps (see Fig. 8):

(1) Reduce the facial image to 224×224 pixels.
(2) Use a pre-trained VGG-19 (on facial data) to extract the features as the output of the last layer in the network. This step provides 1000 features.
(3) Use a spatial averaging filter to convert this 1000 length vector to a 250 length vector.

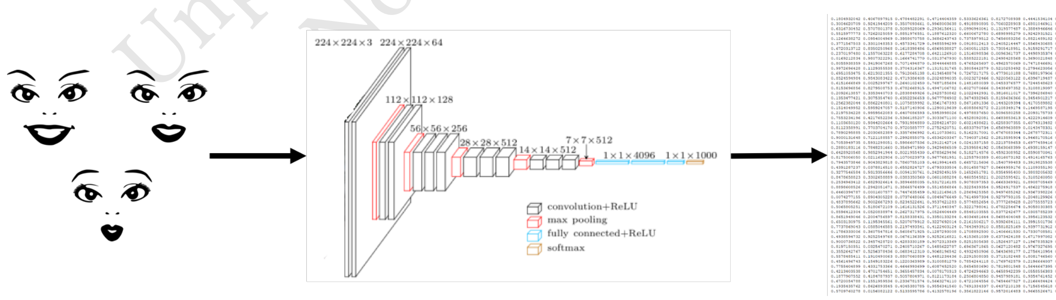

Fig. 8. Process to obtain the facial features using the deep neural network.

[5] Deep features are too many to visualize, and plotting them in the same way as the rest of the features would not convey any meaningful information.

Table 2. Summary of the selected features.

Physiological Features	
Value histogram	Mean, median, SD, skewness, kurtosis of the values.
Spectral histogram	Mean, median, SD, skewness, kurtosis of the dominant frequency components.
ARMA	Auto-regressive moving average: maps the current value to the history of time series.
GARCH	Generalized Auto-regressive conditional heteroskedasticity: maps the current variance to the historical variance of time series and the heterogeneity of the appearance of the values.
LPC	Linear predictive coding: captures the information about the enveloping shape of the signal.
LFSC	Linear Frequency Spectral coding: LPC in frequency domain.
Facial Features	
Action units	Defines the specific area of the face of the user such as, eyebrows, eyes, nose, lips, chin.
Deep Features	Features extracted from a convolutional neural network.

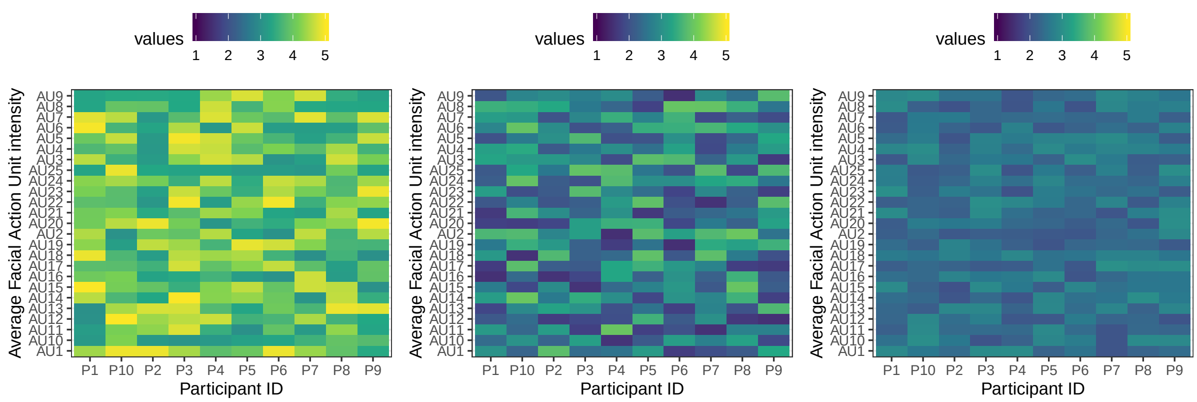

Fig. 9. The average intensity of the facial action units detected for ten random participants in the PM (left), AA (center), and DB (right) studies.

4.4 Feature Selection

One of the techniques to reduce the number of features is to select the most appropriate features, and use them for the training-testing purposes. We use two different feature selection techniques: one linear (Least absolute Shrinkage and Selection Operator—LASSO), and one non-linear (Random Forest [54, 81, 127]). The reason for using LASSO is the fact that for the majority of the pipelines, the number of examples is smaller than the number of features, which is the ideal use-case for LASSO [46, 134]. Furthermore, we decided to also use non-linear feature selection, since there are indications of non-linear relation between the physiological data and the measured behaviour / outcome—cognitive performance in our case [43, 102].

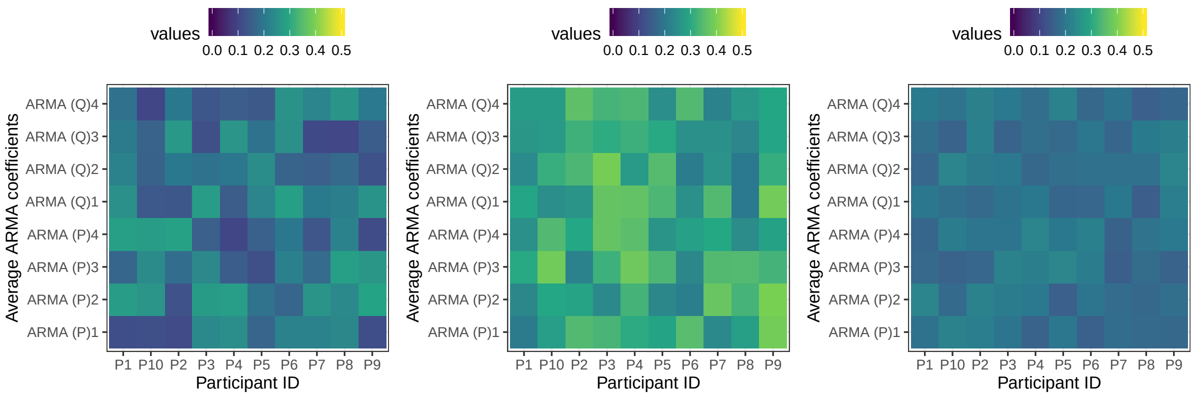

Fig. 10. The average values of ARMA(P=4, Q=4) coefficients for ten random participants in the PM (left), AA (center), and DB (right) studies.

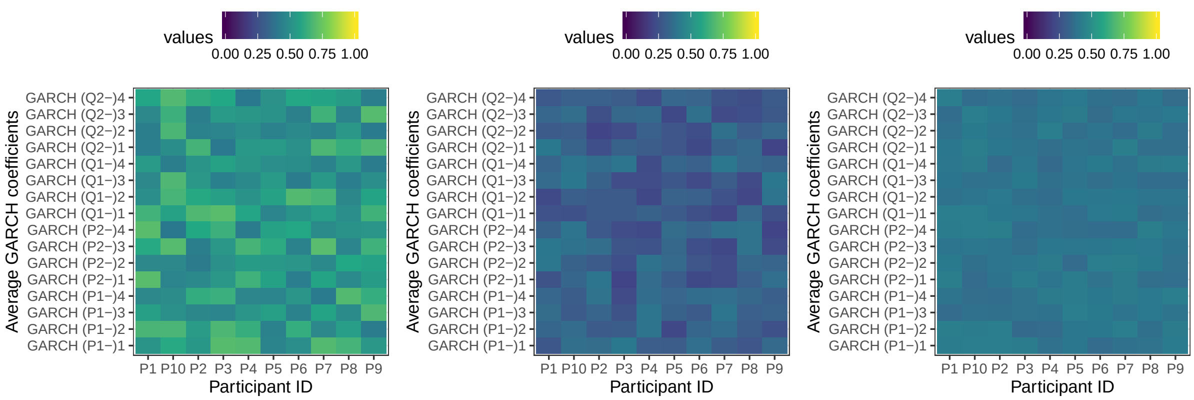

Fig. 11. The average values of GARCH(P1=4, Q1=4, P2=4, Q2=4) coefficients for ten random participants in the PM (left), AA (center), and DB (right) studies.

4.5 Dimensionality Reduction

Apart from feature selection, another way to reduce the number of features is to map the current feature space to a lower dimensional feature space, and conduct the training-testing in the new space. We use two different feature selection techniques: one linear (Principle Component Analysis—PCA), and one non-linear (Kernel PCA [59]). Similar to feature selection, the reason for using a non-linear dimensionality reduction is an indication of non-linear relation between the physiological data and the measured behaviour/outcome—cognitive performance in our case [43, 102]. Another reason for using the non-linear dimensionality reduction technique is that it has been shown to provide better results than the linear techniques [18, 115].

4.6 Prediction: Ensemble Learning

Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance. In this paper, we combine predictions from 7 different algorithms: Support Vector Machines [21]

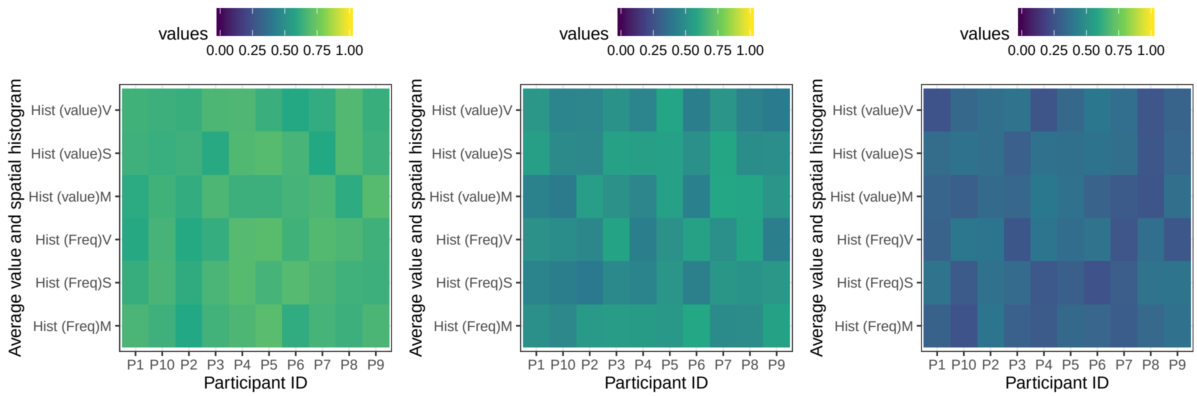

Fig. 12. The average values of value (M:mean, V:variance, S:skewness) and spectral (M:mean, V:variance, S:skewness) histograms for ten random participants in the PM (left), AA (center), and DB (right) studies.

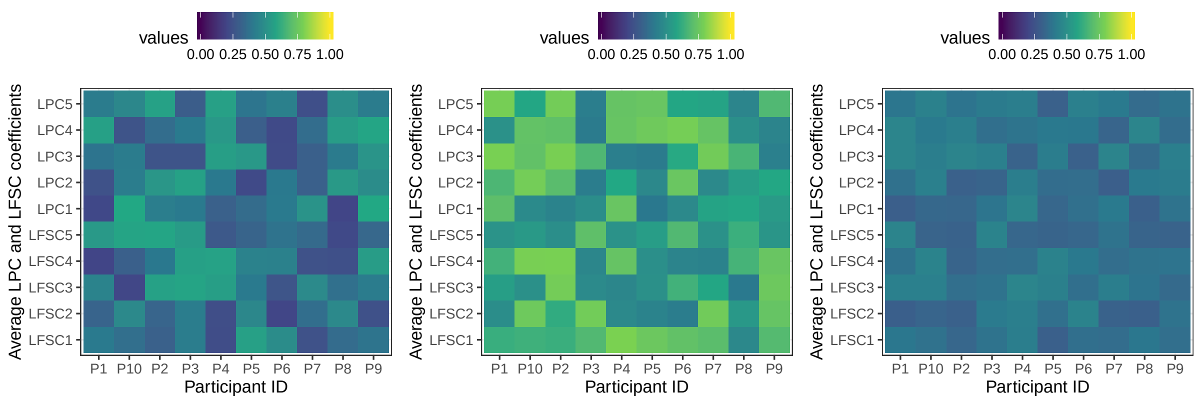

Fig. 13. The average values of LPC (n=5) and LFSC (n=5) coefficients for ten random participants in the PM (left), AA (center), and DB (right) studies.

with linear, radial and polynomial kernels; Gaussian process models [143] with linear, radial and polynomial kernels; and M5 model trees. These methods are designed to improve the stability and the accuracy of Machine Learning algorithms. One way of using the results from multiple models is to use a weighted average from all the prediction algorithms. The weights for individual prediction are considered based on their accuracy during the validation phase. There are 3 major advantages of these methods [8, 43, 100]:

(1) We can compare the performance of the ensemble methods to the diversification of our models predicting cognitive performance. It is advised to keep a diverse set of models to reduce the variability in the prediction and hence, to minimize the error rate. Similarly, the ensemble of models will yield better performance on the test case scenarios (unseen data), as compared to the individual models in most of the cases.
(2) The aggregate result of multiple models always involves less noise than the individual models. This leads to model stability and robustness.

(3) Ensemble models can be used to capture the linear, as well as the non-linear relationships in the data. This can be accomplished by using two different models and forming an ensemble of the two.

4.7 Training, Validation, and Testing Setup

Initially, we perform **out-of-sampling testing** (i.e., leave-one-participant-out), dividing all 3 first datasets into 3 subsets: (1) training, (2) validation, and (3) testing. We keep the testing set aside (10 % from each study). The datasets are split based on participant identifiers. All the models are trained and validated using the training and validation sets with a cross validation. The cross-validation is performed using leave-one-participant-out. In Table 3, pipelines with IDs 1, 4, 9, and 13 are examples of "pure" out-of-sampling testing, where we used the same dataset(s) both for training and testing. In the next stage, we perform **out-of-study testing** (i.e., leave-one-task-out)—that is training on entirely different dataset(s)—and thus context(s)—from the one(s) on which we are testing. This was intended to unveil features that assess cognitive performance reliably across different contexts (i.e., engineering generalizable features). In Table 3, pipelines with IDs 10–12 reflect exactly what we mean by "out-of-study testing," by using 2 study datasets for training, and a 3rd different study dataset for testing. All pipelines were compared based on the Normalized Root Mean Squared Error (NRMSE). The Root Mean Squared Error (RMSE) is calculated using the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{Number\ of\ samples}(predicted_i - original_i)^2}{Number\ of\ samples}} \tag{7}$$

Once we have calculated the RMSE, we normalise it to obtain NRMSE using the following formula:

$$NRMSE(\%) = 100 \times \frac{RMSE}{original_{max} - original_{min}} \tag{8}$$

NRMSE is the proposed metric for student models [98], and is used widely in learning technologies [84] for measuring the accuracy of learning prediction. Another reason for using NRMSE is that it penalizes the larger errors (since the errors are squared before addition), thus making NRMSE a high-quality metric for evaluating predictions. The pipelines were also compared based on the R-statistic measure describing feature generalizability, as we explain in the next section.

4.8 Feature Generalizability Index (FGI)

To measure the generalizability of the features, we examine whether the NRMSE values from the cross-validation and the testing (i.e., out-of-sampling or out-of-study) phases are similar. To this end, we require a statistical test to show the similarity between the two distributions. Since there is no theoretical distribution characterising about the NRMSE values, we require a non-parametric test for checking the similarity of two populations [23]. The ANOSIM (ANalysis Of SIMiliarity) test is non-parametric and bears the null-hypothesis that the two (or more) groups compared have a different mean and variance [23]. Thus, by rejecting the null-hypothesis, one can deduce the similarity of the two NRMSE distributions—in our case: one from the cross-validation and the other from the testing (i.e., out-of-sampling or out-of-study).

Once we have completed all the steps in the pipeline setup, we obtain a list of training (cross-validation) and testing NRMSE per user. The generalizability index of the top features will be the effect size of an ANOSIM. To test for the generalizability (i.e., to conduct ANOSIM) of a given feature set, we require that training and testing datasets come from different studies. Otherwise, the testing NRMSEs are supposed to be similar. Thus, we do not perform this procedure in pipelines with the same training and testing datasets, such as pipelines with IDs 1, 4, 9, and 13 (see Table 3). In cases where the ANOSIM test yields a significant result, the feature set will be considered "generalizable". The R-statistic from the ANOSIM is calculated as follows:

$$R = \frac{mean\ ranks\ between\ groups\ -\ mean\ ranks\ within\ groups}{N(N-1)/4} \tag{9}$$

The denominator ensures that the value of R is between $+1$ and -1, with 0 designating a complete random grouping. The statistical significance of the observed R is assessed by permuting the grouping vector to obtain the empirical distribution of R under null-model.

4.9 Benchmarking the Generalizable Features

After establishing a generalizability metric (FGI), and NRMSE baselines against which to compare (see Table 3), we used the independent dataset from Study 4 (Gaze-based game) to bench-mark the reliability of the generalizable features we previously engineered. We process the new dataset using the same methods we used for the other 3 studies, as previously described. However, we only compute the features for the pipelines that were found to be generalizable (IDs 10–12, see Table 3), and applying the methods described in this section. For comparison, we also compute the features for those pipelines that were shown to be context-specific, or else "non-generalizable" (IDs 1, 5, 9, and 13, see Table 3). Once we compute the features, we use the same ensemble prediction algorithms to predict participants' cognitive performance in Study 4 (Gaze-based game). Then, we run a series of pairwise Wilcoxon signed-rank tests to compare the NRMSE of the generalizable features vs. the non-generalizable features. We use a non-parametric test, since there is no empirical or theoretical basis for assuming any known statistical distribution for the NRMSE values.

5 RESULTS AND DISCUSSION

We test a total of 156 pipelines assembled by 3 data type combinations, 4 feature selection **or** dimensionality reduction techniques, and 13 cross-training and cross-testing combinations. Table 3 summarizes the results from the top 13 most accurate pipelines in predicting cognitive performance (one for each training-testing combination). For brevity, the pipelines are assigned with a numerical ID (i.e., 1^{st} column of Table 3). Pipelines with IDs: 1, 5, and 9 are those in which the training and testing datasets came from the same study (i.e., self-training-testing with out-of-sampling testing). IDs: 1–9 are the pipelines resulting from combinations with one dataset used for training, and one dataset used for testing (i.e., single training-testing, and either out-of-sampling or out-of-study testing). IDs: 10–12 are the pipelines resulting from combinations with two datasets used for training, and one dataset used for testing (i.e., out-of-study testing). For example, ID: 1 is the pipeline with the best NRMSE score of 10.29 % (SD = 2.5 %) when using the dataset from the Pac-Man (PM) study for both training and testing. The corresponding features for ID: 1 are FFT, value and spectral histograms from physiological data, and AU for facial data, selected with the LASSO feature selection technique. The feature generalizability index could not be computed here because the training and testing datasets are the same. **The random baselines for the performance prediction for PM, AA and DB are 44.51, 32.93 and 47.25, respectively. Hence, we observe that the resulting NRMSEs of the 13 most accurate pipelines outperform the random baseline in all the studies** (see Table 3 and Figure 14). The random baselines were calculated using the same distribution as the scores from the individual studies, and by creating random distributions based on the statistics of the normalized scores.

5.1 Selecting Generalizable Features

Table 3 shows the NRMSE values for all the training and testing pairs. As expected, single cross-training-testing (IDs: 2, 3, 4, 6, 7, 8) yields worse prediction than the self-training-testing (IDs: 1, 5, 9). Moreover, we observe that the best feature selection (or dimensionality reduction) method for the single cross training-testing (IDs: 2, 3, 4, 6, 7, 8) is Random Forest (RF). Instead, the best feature selection (or dimensionality reduction) method for the self training-testing (IDs: 1, 5, 9) is LASSO. Interestingly, when we use two datasets for training in cross-training-testing (IDs: 10, 11, 12), we achieve similar prediction results to self-training-testing (IDs: 1, 5, 9). We observe that the best feature selection (or dimensionality reduction) method for these cases (IDs: 10, 11, 12) is Kernel PCA. When we merge all 3 datasets together and perform a simple training-testing approach, we attain the best prediction results (ID: 13). In the case of merged training-testing, the best feature selection (or dimensionality reduction) method is the Random Forest. In the remainder of this section, we will discuss the top features in predicting cognitive performance from a data type perspective (physiological and facial).

Finding #1: The best feature selection technique for the same training and testing context is **LASSO**.
Finding #2: The best feature selection technique for training on data from one context and testing on another is **Random Forest (RF)**.
Finding #3: The best dimensionality reduction technique for training and testing in multiple contexts is **Kernel PCA**.

5.2 Engineering Generalizable Physiological Features

In Table 3, we observe a distinction between physiological features that are context-specific, and those that are generalizable. On one hand, we can see from the single training-testing (IDs: 1–9) that for self-training-testing (IDs: 1, 5, 9) the most important features are the FFT and histograms (ID: 1), FFT and LPC (ID: 5), and histograms for EDA and BVP in particular (ID: 9). However, when using the FFT, LPC, LFSC and value histograms in single cross-training-testing (IDs: 2, 3, 4, 6, 7, 8), we obtain a high prediction error. Thus, these features do not generalize to other contexts. This lack of generalizability, and the high prediction error, indicate context-specific features. On the other hand, the most important features from the multi-dataset cross training-testing (IDs: 10–12), are the feature sets of GARCH and spectral histogram. The fact that we achieve low error rates in the pipelines with ID: 10–12, indicates that these feature sets are generalizable and context-agnostic. Moreover, GARCH and ARMA feature sets emerge among the most important ones when we merge the three datasets and perform regular training-testing. This is yet another indication that GARCH and ARMA feature sets do not depend on context. These findings demonstrate that we were able to produce generalizable features from data of physiological responses to accurately predict cognitive performance in a diverse set of contexts.
Finding #4: The most **generalizable** physiological features are **GARCH** and **spectral histogram**.
Finding #5: The most **context-specific** physiological features are **FFT**, **value and spectral histogram**.

5.3 Engineering Generalizable Facial Features

Similarly, in Table 3 we also note a clear distinction forming between facial features that are context-specific, and facial features that are generalizable. Action Units (AUs) emerge as the most accurate features in assessing cognitive performance both in single training-testing (IDs: 1–9) and in self-training-testing (IDs: 1, 5, 9). In other words, using the AUs to test on the same dataset with the one used for training, yields a low prediction error. However, the AUs do not generalize well to contexts outside which they were trained (IDs: 2, 3, 4, 6, 7, 8). Thus, the lack of generalizability that AUs display, combined with their low prediction error when the same context is used for both training and testing, renders AUs a context-specific feature in predicting cognitive performance. On the contrary, when it comes to multi-dataset cross-training-testing (IDs: 10–12), we observe that the deep features emerge as the most important feature set. The fact that we achieve low error rates in the models with ID 10–12, suggests that the deep features are a generalizable, context-agnostic feature set. Deep features are also among the most important feature sets when we merge the three datasets and perform regular training-testing. This is yet another indication that deep feature sets do not depend on context. These findings demonstrate that we were able to produce generalizable features from data about facial expressions that accurately predict cognitive

performance in a diverse set of contexts.
Finding #6: The most **generalizable** facial features are **deep features**.
Finding #7: The most **context-specific** facial features are **Action Units (AUs)**.

5.4 On Feature Generalizability

To evaluate the capacity of our features to reliably assess cognitive performance in diverse contexts, we introduce a new measure—the feature generalizability index (FGI). We computed the FGI, as described in Section 4.8, for each pipeline using the *R*-statistic. The *R*-statistic designates how generalizable the pipeline is, and thus reveals which is the most important feature set. A non-significant *R*-statistic in the Table 3 shows that there is a considerable amount of contextual information in the pipeline, which leads to a different testing NRMSE (IDs: 2, 3, 5, 6, 7, 8). Conversely, a significant *R*-statistic shows that the NRMSE scores, produced from cross-validation testing, are similar and thus the pipelines generalise from the training set to the testing set (IDs: 10, 11, 12). We observe that the testing NRMSE scores of the generalizable pipelines (IDs: 10, 11, 12) appear relatively similar to pipeline ID: 13, where we have merged the 3 datasets from the 3 studies, and perform regular training-testing. All in all, we were able to quantify how generalizable the features produced from physiological responses and facial expressions are in reliably predicting cognitive performance in diverse contexts.
Finding #8: FGI measures the **generalizability of features** that assess **cognitive performance** stemming from **physiological responses** and **facial expressions**.

Table 3. Best pipelines identified by their IDs corresponding to 13 cross-training and cross-testing combinations for Pac-Man (PM), Adaptive Assessment (AA), and Debugging (DB) datasets. The data types ("Both" for physiological & facial data) and the selection/reduction technique are displayed next. Accuracy in predicting cognitive performance is presented next, described by minimizing Normalized Root Mean Squared Error (NRMSE) in %, and feature generalizability index (R) in a -1 to +1 scale, followed by the selected feature sets. We observe that feature sets of pipelines 10, 11, and 12 display the best NRMSE and R index, respectively. **The random baselines for the PM, AA, and DB are 44.51, 32.93, and 47.25, respectively.** When no particular physiological data type is mentioned (e.g., EDA), the entirety of physiological data was included in the prediction.

ID	Training	Testing	NRMSE (SD)	Data (technique)	R (p)	Selected Feature Set
1	PM	PM	10.29 (2.5)	Both (LASSO)	N/A	**E4:** FFT, value and spectral histograms **Face:** AUs
2	PM	AA	18.46 (3.2)	Both (RF)	-0.007 (> 0.05)	**E4:** FFT, (BVP, HR, EDA) LPC **Face:** AUs
3	PM	DB	19.67 (3.8)	Both (RF)	-0.012 (> 0.05)	**E4:** FFT, (BVP, HR, EDA) LPC **Face:** AUs
4	AA	PM	19.32 (3.1)	Both (RF)	N/A	**E4:** LPC, LFSC, value histograms **Face:** AUs
5	AA	AA	10.77 (2.4)	Both (LASSO)	-0.005 (> 0.05)	**E4:** LPC, FFT **Face:** AUs
6	AA	DB	15.30 (3.9)	Both (RF)	-0.06 (> 0.05)	**E4:** LPC, LFSC, value histograms **Face:** AUs
7	DB	PM	19.37 (3.1)	Both (RF)	0.04 (> 0.05)	**E4:** LPC, LFSC, value histograms **Face:** AUs
8	DB	AA	15.75 (3.8)	Both (RF)	0.07 (> 0.05)	**E4:** LPC, LFSC, value histograms **Face:** AUs
9	DB	DB	11.15 (2.3)	Both (LASSO)	N/A	**E4:** (EDA, BVP) value and spectral histograms **Face:** AUs
10	**PM, AA**	**DB**	**9.24 (1.6)**	**Both (Kernel PCA)**	**0.17 (< 0.05)**	**E4: GARCH, spectral histogram** **Face: deep features**
11	**PM, DB**	**AA**	**8.27 (2.1)**	**Both (Kernel PCA)**	**0.32 (< 0.01)**	**E4: GARCH, spectral histogram** **Face: deep features**
12	**AA, DB**	**PM**	**8.26 (1.9)**	**Both (Kernel PCA)**	**0.35 (< 0.01)**	**E4: GARCH, spectral histogram** **Face: deep features**
13	PM, AA, DB	PM, AA, DB	8.17 (1.6)	Both (RF)	N/A	**E4:** GARCH, ARMA **Face:** deep features

5.5 Bench-mark Results for Generalizable Features

As described in Section 4.9, we use the entirely independent dataset of Study 4 (Gaze-based game) to evaluate the accuracy of the best-performing features in assessing cognitive performance, comparing between context-specific and context-agnostic (generalizable) feature engineering approaches. Table 4, illustrates the pipelines we use for comparison, with IDs: 1, 5, 9, and 13 falling into the context-specific category, and IDs: 10–12, falling into the context-agnostic category. In Table 4, we observe that the NRMSE values for context-specific features (IDs: 1, 5, and 9) are higher than those for context-agnostic (generalizable) features (IDs: 10–12), produced with out-of-study testing. However, to reliably support this claim, we run pairwise comparisons using Wilcoxon signed-rank tests among the NRMSE values for all selected pipelines shown in Table 4. Overall, the results show that the pipelines using generalizable features (IDs: 10–12) perform significantly better than the pipelines using context-specific features (IDs: 1, 5, and 9) in reliably predicting cognitive performance in Study 4 (gaze-based game), as shown in Table 5. Notably, pipeline 13 has the lowest NRMSE value, since it is trained on all 3 previous datasets (i.e., Pac-Man game, adaptive-assessment learning, and code-debugging).

We emphasize that the evaluation of the generalizability of the engineered features is conducted in a context that is highly representative of Ubiquitous Computing scenarios. Not only does Study 4 involve gaze-based interactions, but the entire sample population consists of school students aged 8–14 years, in contrast to all 3 previous studies with participants aged 17–49, 18–21, and 20–22 years, for Pac-Man game, adaptive-assessment learning, and code-debugging, respectively.

Finding #9: Generalizable features reliably **assess cognitive performance** in **diverse contexts**, across **different tasks**, and with **diverse sample populations**.

Table 4. Evaluation of generalizable features (ID: 10–12) and non-generalizable features (ID: 1,5,9). Each pipeline is evaluated in terms of NRMSE values from the ensemble prediction. The whole dataset from the Study 4 is used for testing. The random baseline for the performance in Study 4 is 34.65.

Pipeline ID from Table 3	Trained on	Selected Feature Set	NRMSE (SD)
1	PM	**E4**: FFT, value and spectral histograms **Face**: AUs	15.67 (3.20)
5	AA	**E4**: LPC, FFT **Face**: AUs	15.72 (3.23)
9	DB	**E4**: (EDA, BVP) value and spectral histograms **Face**:AUs	17.88 (3.73)
10	PM, AA	**E4**: GARCH, spectral histogram **Face**: deep features	9.76 (2.89)
11	PM,DB	**E4**: GARCH, spectral histogram **Face**: deep features	9.39 (2.71)
12	AA, DB	**E4**: GARCH, spectral histogram **Face**: deep features	9.16 (2.41)
13	PM, AA, DB	**E4**: GARCH, ARMA **Face**: deep features	8.64 (1.93)

Table 5. The pairwise comparisons between the NRMSE values from the pipelines using the generalizable features (ID: 10–12) and the context-specific features (ID: 1,5,9, and 13). The entire dataset from Study 4 (gaze-based game) was used to test the ability of both categories features to predict cognitive performance. The values in the cells are the Wilcoxon test-statistic and the corresponding p-values in the parentheses.

Pipeline ID	**1**	**5**	**9**	**10**	**11**	**12**	**13**
1	-	424 (0.60)	268 (0.04)	683 (0.0001)	769 (0.0001)	768 (0.0001)	761 (0.0001)
5	-	-	241 (0.01)	658 (0.0001)	741 (0.0001)	742 (0.0001)	734 (0.0001)
9	-	-	-	719 (0.0001)	781 (0.0001)	784 (0.0001)	784 (0.0001)
10	-	-	-	-	536 (0.01)	532 (0.01)	522 (0.01)
11	-	-	-	-	-	380 (0.85)	389 (0.96)
12	-	-	-	-	-	-	386 (0.92)
13	-	-	-	-	-	-	-

5.6 Context-specific Features

For the self-training-testing pipelines (IDs: 1, 5, 9), we note that the variable importance (from the random forest) reflects the context-sensitivity as well. Further inspection of the most important features reveals the following feature sets for the three studies:

(1) **PM**: *Action Units from facial features*→cheek raiser, lip corner puller, upper lip raiser, lip corner depressor, lip stretcher.
Physiological features→Most dominant frequency HR (FFT-1), mean and variance for HR and EDA.
(2) **AA**: *Action Units from facial features*→inner brow raiser, outer brow raiser, nose wrinkler, dimpler, lip tightener.
Physiological features→first LPC coefficient HR, mean, and variance for HR and BVP.
(3) **DB**: *Action Units from facial features*→brow lowerer, lid tightener, upper lid raiser, chin raiser, lip suck.
Physiological features→mean frequency HR, mean and variance for BVP, and EDA.

We observe that the most important set of facial features from the three studies have almost no overlap across all three studies, while the most important set of physiological features display low overlap when it comes to the histogram-based features. This proves the fact that self-training-testing produces context-specific features, since the training is done on one dataset only.
Finding #10: There is a substantial amount of context-specific information (**variability across contexts**) in the physiological (FFT, LPC and histogram based features) and facial data (Action Units).

5.7 Implications

Our results show that there are two sets of features, one from physiological data and one from facial data, that yield the highest FGI (ID: 10, 11, 12 in Table 3). For the physiological data, these are the coefficients from the GARCH model and the features computed from the spectral histogram (mean, SD, skewness, kurtosis and maximum), whereas for the facial data, it is the deep features (computed by a pre-trained deep neural network). Instead, context-specific features include the FFT, value histograms, and LPC coefficients for the physiological

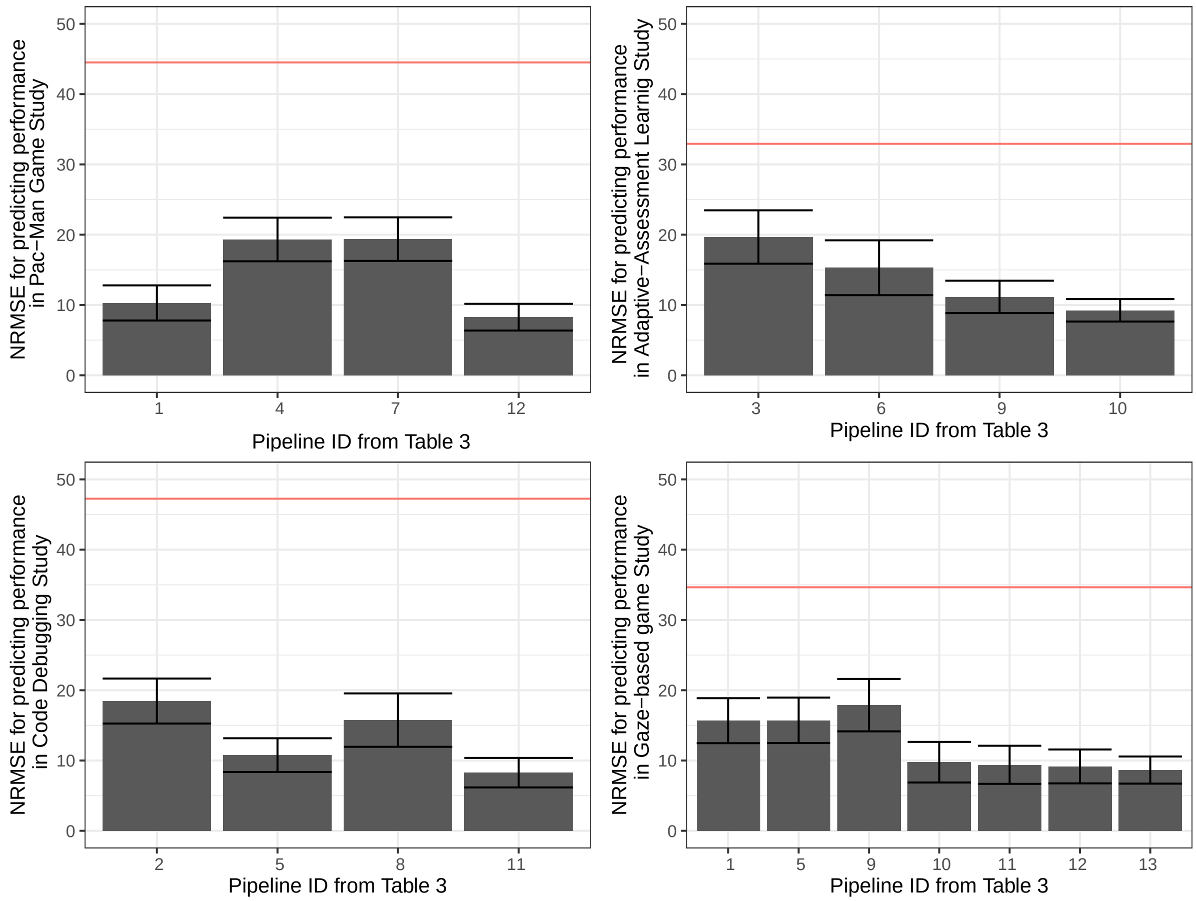

Fig. 14. Comparison of the 13 best pipelines, grouped by the study in which they were tested. The error bars represent the standard deviation, and the red lines represent the random NRMSE baseline.

data, and the Action Units (AUs) computed from the facial data (see Table 3 ID: 1, 5, 9). This designates that in diverse contexts and across different tasks, there can be two kinds of features: (a) those that generalize to diverse contexts (context-agnostic), and (b) those specific to the target context (context-specific).

We observe that GARCH features from physiological data emerged as one of the most generalizable feature set. This indicates that modelling the variability of physiological timeseries produces generalizable features across diverse contexts. GARCH models have a number of advantages over contemporary time series modelling methods. For example, GARCH does not require any prior quantization (as opposed to Markov chain based methods), since it is an approach designed for continuous time-series data. Plus, the length of history used by GARCH can be empirically decided by a likelihood estimation, and there is no need for contingency counts, as opposed to N-gram based methods. Also, GARCH describes the "conditional variance" in the time series, as opposed to classical modelling of "conditional mean" (auto-regression). These properties of GARCH models render them an efficient time-series modelling technique [19, 32, 75]. In fact, the aforementioned properties of GARCH may be the reason why GARCH model-based features achieve high FGI.

The fact that "deep features" from facial data emerged as the most generalizable, while the AUs turned out to be context-specific, speaks to the context-sensitivity of emotions. In fact, AUs are typically used for gauging emotions through facial expressions [29, 31, 49, 50]. Our findings indicate that deep features may be one way to obtain generalizable features. So far, deep neural networks have been utilized in "transfer learning", where part of the model is transferred between different but related contexts in the domains of energy [62, 101], linguistics [57, 63], and image processing [22, 151]. In our case, we used a relatively simple, pre-trained deep neural network (VGG-19) to extract the features from facial expressions manifested in 3 different study contexts. By cross training-testing, we opted for generalizabilty, showing that when it comes to facial data, deep features capture more intricacies than the AUs do.

As previous findings suggest, cognitive performance relies on the state of the many underlying cognitive processes [140], and it is affected by a plethora of factors [2, 5, 20, 58, 60, 79, 108, 137, 150]. Depending on the context and the task at hand, different cognitive processes may manifest. Thus from the outset, accurately gauging cognitive performance is not an easy feat. Multiple instances in literature have utilized physiological responses and facial expressions in monitoring cognitive performance [9] for increasing productivity [83, 105], deciding when one can be interrupted [48, 114], monitoring workload [72, 122], gauging enjoyment [65, 130, 135], and facilitating learning [10, 28, 29, 33, 35, 36, 42, 49].

Across all these instances of prior work, one can quickly notice the diversity of the contexts in which some aspect of cognitive performance was measured. Inevitably, instances such as the above, are almost always tailored to measure aspects of cognitive performance with great accuracy, but within a strictly specific context and during a specific task at hand. Thus, when it comes to assessing cognitive performance in a new context, little if any knowledge can be transferred, and prediction models have to be generated again through exhaustive trial-and-error approaches. Although the need for generalizability in ML has been stressed before in multiple instances, such as music information retrieval [112], personality assessment [14], predicting driver intentions [99], and developing BCIs [91], little progress has been made towards developing generalizable features.

Instead, most ML approaches that claim generalizability, focus on "transfer learning" in deep learning [136]. However, more recently transfer learning (TL) typically assumes deep learning, since the computational power has become on par with computational needs. Moreover, TL requires an already trained model, parts of it, or a model trained on related data (e.g., recognizing cats) that is introduced to a new but related context for completing a similar task (e.g., recognizing objects). Thus, TL is fundamentally different from engineering generalizable features. Recent work by Hutt et al., on producing generalizable affect detection from usage analytics of online learning platforms, is perhaps an instance that approximates our work the most [64]. Even so, their selected features were generic and "hand-picked," while relying on extraordinary big sample sizes ($> 69,000$ users). In our work, we attempt to overcome the lack of generalizability that characterizes most of the ML approaches in literature, by introducing an ML methodology for engineering generalizable features in a systematic and near-automatic fashion, with data from attainable sample sizes.

By drawing on 4 datasets from 4 independent studies, we were able to engineer features that generalize well for assessing cognitive performance. Engineering features that reliably assess cognitive performance in diverse contexts yields novel opportunities, not only in the realm of Ubiquitous Computing and HCI, but also in Cognitive Psychology and Neuroergonomics. Indeed, generalizable features, in combination with the ubiquity of wearable and image-capture devices, enable the around-the-clock monitoring of the states of our cognitive processes. This could bear tangible benefits in domains such as the ones mentioned earlier (e.g., increasing productivity, facilitating learning, etc.), and could also be incorporated in existing architectures for delivering improved wearable cognitive assistance [52, 88, 139], and eventually pave the way for cognition-aware systems [17, 30, 89].

But perhaps the most important contribution of this work lies in the methodology applied within, and in the generalizable knowledge to which it has contributed. Besides highlighting which physiological and facial features one can use to reliably assess cognitive performance in diverse contexts, the methodology per se can be applied entirely outside the realm of cognitive performance. Thus, in this work we have contributed towards the creation of knowledge that is more abstract than particular instances, leading to generalized theories [61].

5.8 Limitations

Besides the applicable findings and the exciting methodological potential this work bears, it also comes with significant limitations that we need to address. First, we postulate that the 4 studies (and the 4 corresponding datasets), on which this work draws, encompass the major portion of the cognitive processes that underpin human cognition. Although we do not expect that all cognitive processes manifested to the same extent across all studies and all participants, the selected study contexts (i.e., Pac-Man game, adaptive-assessment learning, code-debugging, and a gaze-based game) required different levels of decision-making, problem-solving, memory recall, learning, and of course attention. In Study 4 (gaze-based game), we were surprised to discover that our generalizable features performed considerably well in predicting the cognitive performance of school students. So far, we had engineered our features entirely based on datasets collected from adults performing a variety of cognitive tasks. We did however try to control as many variables as possible by applying the same experimental protocol across all 4 studies (see Fig. 1).

Next, this work assumes that cognitive performance can be characterized by the score that one achieves in a mental task, and can be reflected in one's physiological responses and facial expressions. On one hand, our approach is by design computational, and thus it relies a priori on quantified and objective measures of performance such as scores. On the other hand, there is an amassing body of evidence on the connection of physiological responses and facial expressions with cognitive performance [26, 56, 149]. In this work, we did not consider physiological responses measured by EEG and eye-tracking, simply due to requiring stationary settings—our intention is to move outside the lab. Having said that, we need to acknowledge that in this stage, this work builds on studies that have taken place entirely in control settings. In this way, we were able to minimize most of the confounding factors that impact physiological responses (e.g., movement), and ensure that the proposed methodology yields the desirable results before we transfer it outside the lab.

Finally, we do recognize the fact that we have not deployed any means for directly collecting feedback on the cognitive workload our participants exhibited. For example, administering a NASA-TLX questionnaire [55] would have shed light on the cognitive workload our participants experienced when completing a cognitive task through self-assessment. In turn, utilizing self-reported (cognitive) workload could have enabled us to estimate cognitive performance in an even more accurate, and perhaps more generalizable fashion. Thus, purely relying on scores bears the drawback of potentially miss-classifying high-performing individuals, who may exhibit little or no physiological and facial expression effects, due to reduced effort invested on their part. However, our assumption here is that high-performing individuals, who do not experience any physiological effects due to cognitive workload, are outliers. Detecting and modeling such outliers would require more sophisticated approaches, such as the Extreme Value Theorem and Copula Theory [86], and are currently outside the scope of this work.

6 CONCLUSION AND FUTURE WORK

In this work, we introduce a machine learning methodology for engineering generalizable features from physiological responses and facial expressions that assess cognitive performance. Our methodology draws on 4 independent studies, that followed a highly-similar experimental protocol, and 4 corresponding datasets from a total of 123 participants, exhibiting varying levels of problem-solving, decision-making, and learning processes during the completion of the tasks at hand. Our results show that LASSO is the best feature selection technique when it comes to training and testing in the same context, whereas Random Forest performs better when it comes to testing in one context and training in another. Kernel PCA emerged as the best dimensionality reduction technique for training and testing in multiple contexts.

Our methodology revealed that the most generalizable features in reliably assessing cognitive performance are GARCH with spectral histogram and deep features from data of physiological responses and facial expressions, respectively. On the contrary, the most context-specific features are FFT, value and spectral histograms for physiological responses, and Action Units for facial expressions. By introducing a feature generalizability index (FGI), we showcase how our methodology can be applied for engineering generalizable features outside the realm of cognitive performance.

As for future work, we plan to extend our methodology to consider mobility and physical activity by supplying it with the corresponding data streams (e.g., accelerometer values) for measuring cognitive performance outside the lab. We also plan to use our technique with additional data that reveal cognitive performance such as facial thermal imaging [1]. Finally, we plan to explore how generalizable our approach can be in assessing cognitive performance during collaborative tasks, using the features engineered in this work.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support of the Swiss National Science Foundation (SNSF) under grant number: 184146.

REFERENCES

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 33.

[2] Phillip L Ackerman. 2011. *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications.* American Psychological Association.

[3] Mary Ainley, Matthew Corrigan, and Nicholas Richardson. 2005. Students, tasks and emotions: Identifying the contribution of emotions to students' reading of popular culture and popular science texts. *Learning and Instruction* 15, 5 (2005), 433–447.

[4] Carol Alexander. 2001. *Market models: A guide to financial data analysis.* John Wiley & Sons.

[5] Paula Alhola and Päivi Polo-Kantola. 2007. Sleep deprivation: Impact on cognitive performance. *Neuropsychiatric disease and treatment* (2007).

[6] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6 (2016).

[7] B Atal and M Schroeder. 1978. Predictive coding of speech signals and subjective error criteria. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'78.*, Vol. 3. IEEE, 573–576.

[8] Ran Avnimelech and Nathan Intrator. 1999. Boosted mixture of experts: an ensemble learning scheme. *Neural computation* 11, 2 (1999), 483–497.

[9] Ashwin Ramesh Babu, Akilesh Rajavenkatanarayanan, James Robert Brady, and Fillia Makedon. 2018. Multimodal approach for cognitive task performance prediction from body postures, facial expressions and EEG signal. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. ACM, 2.

[10] Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

[11] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[12] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. 2005. Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 568–573.

[13] Edison L Bell and Richard Franke. 1976. *A numerical comparison of Toeplitz equation solving algorithms.* Technical Report.

[14] Wiebke Bleidorn and Christopher James Hopwood. 2019. Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review* 23, 2 (2019), 190–203.

[15] Herbert Bless. 2000. The interplay of affect and cognition: The mediating role of general knowledge structures. (2000).

[16] Guy A Boy. 1998. Cognitive function analysis for human-centered automation of safety-critical systems. In *CHI*, Vol. 98. 265–272.

[17] Andreas Bulling and Thorsten O Zander. 2014. Cognition-aware computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83.

[18] Christian Callegari, Lisa Donatini, Stefano Giordano, and Michele Pagano. 2018. Improving stability of PCA-based network anomaly detection by means of kernel-PCA. *International Journal of Computational Science and Engineering* 16, 1 (2018), 9–16.

[19] Joshua CC Chan and Angelia L Grant. 2016. Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Economics* 54 (2016), 182–189.

[20] Yu-Kai Chang, Jeffrey D Labban, Jennifer I Gapin, and Jennifer L Etnier. 2012. The effects of acute exercise on cognitive performance: a meta-analysis. *Brain research* 1453 (2012), 87–101.

[21] Olivier Chapelle and Vladimir Vapnik. 2000. Model selection for support vector machines. In *Advances in neural information processing systems*. 230–236.

[22] Dan C Cireşan, Ueli Meier, and Jürgen Schmidhuber. 2012. Transfer learning for Latin and Chinese characters with deep neural networks. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.

[23] K Robert Clarke. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology* 18, 1 (1993), 117–143.

[24] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. 2007. Observer-based measurement of facial expression with the Facial Action Coding System. *The handbook of emotion elicitation and assessment* (2007), 203–221.

[25] Scotty D Craig, Sidney D'Mello, Amy Witherspoon, and Art Graesser. 2008. Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning. *Cognition and Emotion* 22, 5 (2008), 777–788.

[26] Mihaly Czikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. New York: Harper & Row.

[27] Ryan SJ d Baker, Sujith M Gowda, Michael Wixon, Jessica Kalka, Angela Z Wagner, Aatish Salvi, Vincent Aleven, Gail W Kusbit, Jaclyn Ocumpaugh, and Lisa Rossi. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *International Educational Data Mining Society* (2012).

[28] Ryan SJ d Baker, Gregory R Moore, Angela Z Wagner, Jessica Kalka, Aatish Salvi, Michael Karabinos, Colin A Ashe, and David Yaron. 2011. The dynamics between student affect and behavior occurring outside of educational software. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 14–24.

[29] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 103.

[30] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 47.

[31] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 43.

[32] Chaido Dritsaki. 2017. An empirical evaluation in GARCH volatility modeling: Evidence from the Stockholm stock exchange. *Journal of Mathematical Finance* 7, 02 (2017), 366.

[33] Sidney D'Mello. 2012. Monitoring affective trajectories during complex learning. *Encyclopedia of the Sciences of Learning* (2012), 2325–2328.

[34] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[35] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29 (2014), 153–170.

[36] Sidney D'Mello, Blair Lehman, Jeremiah Sullins, Rosaire Daigle, Rebekah Combs, Kimberly Vogt, Lydia Perkins, and Art Graesser. 2010. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *International conference on intelligent tutoring systems*. Springer, 245–254.

[37] Paul Ekman and Dacher Keltner. 1997. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture* (1997), 27–46.

[38] Robert Engle. 2001. GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of economic perspectives* 15, 4 (2001), 157–168.

[39] K Anders Ericsson, Robert R Hoffman, Aaron Kozbelt, and A Mark Williams. 2018. *The Cambridge handbook of expertise and expert performance*. Cambridge University Press.

[40] Barbara L Fredrickson. 1998. What good are positive emotions? *Review of general psychology* 2, 3 (1998), 300.

[41] John W French, Ruth B Ekstrom, and Leighton A Price. 1963. *Manual for kit of reference tests for cognitive factors (revised 1963)*. Technical Report. Educational Testing Service Princeton NJ.

[42] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 13.

[43] Valeriy V Gavrishchaka, Mark E Koepke, and Olga N Ulyanova. 2010. Boosting-based discovery of multi-component physiological indicators: Applications to express diagnostics and personalized treatment optimization. In *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, 790–799.

[44] Peyvand Ghaderyan and Ataollah Abbasi. 2016. An efficient automatic workload estimation method based on electrodermal activity using pattern classifier combinations. *International Journal of Psychophysiology* 110 (2016), 91–101.

[45] Giuseppe Ghiani, Marco Manca, and Fabio Paternò. 2015. Dynamic user interface adaptation driven by physiological parameters to support learning. In *Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. ACM, 158–163.

[46] Michail N Giannakos, Kshitij Sharma, Ilias O Pappas, Vassilis Kostakos, and Eduardo Velloso. 2019. Multimodal data as a means to understand the learning experience. *International Journal of Information Management* 48 (2019), 108–119.

[47] Martin Gjoreski, Mitja Luštrek, and Veljko Pejović. 2018. My Watch Says I'm Busy: Inferring Cognitive Load with Low-Cost Wearables. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 1234–1240.

[48] Nitesh Goyal and Susan R Fussell. 2017. Intelligent interruption management using electro dermal activity based physiological sensor for collaborative sensemaking. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 52.

[49] Arthur Graesser, Patrick Chipman, Brandon King, Bethany McDaniel, and Sidney D'Mello. 2007. Emotions and learning with auto tutor. *Frontiers in Artificial Intelligence and Applications* 158 (2007), 569.

[50] AC Graesser, Bethany McDaniel, Patrick Chipman, Amy Witherspoon, Sidney D'Mello, and Barry Gholson. 2006. Detection of emotions during learning with AutoTutor. In *Proceedings of the 28th annual meetings of the cognitive science society*. Citeseer, 285–290.

[51] James J Gross. 2002. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology* 39, 3 (2002), 281–291.

[52] Kiryong Ha, Zhuo Chen, Wenlu Hu, Wolfgang Richter, Padmanabhan Pillai, and Mahadev Satyanarayanan. 2014. Towards wearable cognitive assistance. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 68–81.

[53] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 301–310.

[54] Siddharth Hariharan, Dipankar Mandal, Siddhesh Tirodkar, Vineet Kumar, Avik Bhattacharya, and Juan Manuel Lopez-Sanchez. 2018. A novel phenology based feature subset selection technique using random forest for multitemporal PolSAR crop classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11, 11 (2018), 4244–4258.

[55] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[56] Donald Olding Hebb. 1955. Drives and the CNS (conceptual nervous system). *Psychological review* 62, 4 (1955), 243.

[57] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean. 2013. Multilingual acoustic models using distributed deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8619–8623.

[58] G Robert J Hockey. 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological psychology* 45, 1-3 (1997), 73–93.

[59] Heiko Hoffmann. 2007. Kernel PCA for novelty detection. *Pattern recognition* 40, 3 (2007), 863–874.

[60] Eef Hogervorst, Wim Riedel, Asker Jeukendrup, and Jelle Jolles. 1996. Cognitive performance after strenuous physical exercise. *Perceptual and motor skills* 83, 2 (1996), 479–488.

[61] Kristina Höök and Jonas Löwgren. 2012. Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 3 (2012), 23.

[62] Qinghua Hu, Rujia Zhang, and Yucan Zhou. 2016. Transfer learning for short-term wind speed prediction with deep neural networks. *Renewable Energy* 85 (2016), 83–95.

[63] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7304–7308.

[64] Stephen Hutt, Joseph F Grafsgaard, and Sidney K D'Mello. 2019. Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire School Year. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 496.

[65] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. 2018. EngageMon: Multi-Modal Engagement Sensing for Mobile Games. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 13.

[66] Curtis S Ikehara and Martha E Crosby. 2005. Assessing cognitive load with physiological sensors. In *Proceedings of the 38th annual hawaii international conference on system sciences*. IEEE, 295a–295a.

[67] Peter Kabal and Ravi P Ramachandran. 1986. The computation of line spectral frequencies using Chebyshev polynomials. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34, 6 (1986), 1419–1426.

[68] Irvin R Katz and John R Anderson. 1987. Debugging: An analysis of bug-location strategies. *Human-Computer Interaction* 3, 4 (1987), 351–399.

[69] Jeroen Kerkhof, Bertrand Melenberg, and Hans Schumacher. 2010. Model risk and capital reserves. *Journal of Banking & Finance* 34, 1 (2010), 267–279.

[70] Dean G Kilpatrick. 1972. Differential responsiveness of two electrodermal indices to psychological stress and performance of a complex cognitive task. *Psychophysiology* 9, 2 (1972), 218–226.

[71] Victor D Kolesnik, Andrey N Trofimov, Irina E Bocharova, Victor Y Krachkovsky, Boris D Kudryashov, Eugeny P Ovsjannikov, Boris K Trojanovsky, and Sergei I Kovalov. 1997. Method and apparatus for speech compression using multi-mode code excited linear predictive coding. US Patent 5,602,961.

[72] Thomas Kosch, Jakob Karolus, Havy Ha, and Albrecht Schmidt. 2019. Your skin resists: exploring electrodermal activity as workload indicator during manual assembly. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. ACM, 8.

[73] John W Krakauer and Pietro Mazzoni. 2011. Human sensorimotor learning: adaptation, skill, and beyond. *Current opinion in neurobiology* 21, 4 (2011), 636–644.

[74] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.

[75] P Kumar, S Patil, et al. 2016. Volatility Forecasting–A Performance Measure of Garch Techniques With Different Distribution Models. *International Journal of Soft Computing, Mathematics and Control (IJSCMC)* 5, 2/3 (2016).

[76] Peter Lewinski, Tim M den Uyl, and Crystal Butler. 2014. Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics* 7, 4 (2014), 227.

[77] EA Linnenbrink and PR Pintrich. 2003. Motivation, affect, and cognitive processing: What role does affect play. In *annual meeting of the American Educational Research Association, Chicago, IL*.

[78] Marius Matei. 2009. Assessing volatility forecasting models: why GARCH models take the lead. *Romanian Journal of Economic Forecasting* 12, 4 (2009), 42–65.

[79] Joan M McDowd and Fergus IM Craik. 1988. Effects of aging and task difficulty on divided attention performance. *Journal of Experimental Psychology: Human Perception and Performance* 14, 2 (1988), 267.

[80] Ranjana K Mehta and Raja Parasuraman. 2013. Neuroergonomics: a review of applications to physical and cognitive work. *Frontiers in human neuroscience* 7 (2013), 889.

[81] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics* 10, 1 (2009), 213.

[82] Debra K Meyer and Julianne C Turner. 2002. Discovering emotion in classroom motivation research. *Educational psychologist* 37, 2 (2002), 107–114.

[83] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 37.

[84] Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos, Pedro J Muñoz-Merino, and Carlos Delgado Kloos. 2018. Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies* (2018).

[85] Takehiro Moriya, Yutaka Kamamoto, Noboru Harada, Hirokazu Kameoka, and Ryosuke SUGIURA. 2018. Linear predictive coding apparatus, linear predictive decoding apparatus, and methods, programs and recording medium therefor. US Patent App. 15/562,689.

[86] Roger B Nelsen. 2007. *An introduction to copulas*. Springer Science & Business Media.

[87] Roderick I Nicolson and Angela J Fawcett. 2000. Long-term learning in dyslexic children. *European Journal of Cognitive Psychology* 12, 3 (2000), 357–393.

[88] Evangelos Niforatos, Athanasios Vourvopoulos, and Michail Giannakos. [n.d.]. Endowing Head-Mounted Displays with Physiological Sensing for Augmenting Human Learning and Cognition. ([n. d.]).

[89] Evangelos Niforatos, Athanasios Vourvopoulos, and Marc Langheinrich. 2017. Amplifying human cognition: bridging the cognitive gap between human and machine. In *Proceedings of the 2017 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2017 ACM international symposium on wearable computers*. 673–680.

[90] Nargess Nourbakhsh, Yang Wang, Fang Chen, and Rafael A Calvo. 2012. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. ACM, 420–423.

[91] Ewan S Nurse, Philippa J Karoly, David B Grayden, and Dean R Freestone. 2015. A generalizable brain-computer interface (BCI) using machine learning for feature discovery. *PloS one* 10, 6 (2015), e0131328.

[92] Emilio Soria Olivas. 2009. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*. IGI Global.

[93] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[94] Z Papamitsiou and A A Economides. 2013. Towards the alignment of computer-based assessment outcome with learning goals: The LAERS architecture. In *2013 IEEE Conference on e-Learning, e-Management and e-Services*. 13–17. https://doi.org/10.1109/IC3e.2013.6735958

[95] Corinna Peifer, André Schulz, Hartmut Schächinger, Nicola Baumann, and Conny H Antoni. 2014. The relation of flow-experience and physiological arousal under stress—can u shape it? *Journal of Experimental Social Psychology* 53 (2014), 62–69.

[96] Reinhard Pekmn, Thomas Goetz, Wolfram Titz, et al. 2002. Academic emotions in students" self regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist* 37 (2002), 91–106.

[97] Reinhard Pekrun. 2006. The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review* 18, 4 (01 Dec 2006), 315–341. https://doi.org/10.1007/s10648-006-9029-9

[98] Radek Pelánek. 2015. Metrics for evaluation of student models. *Journal of Educational Data Mining* 7, 2 (2015), 1–19.

[99] Derek J Phillips, Tim A Wheeler, and Mykel J Kochenderfer. 2017. Generalizable intention prediction of human drivers at intersections. In *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1665–1670.

[100] Xueheng Qiu, Le Zhang, Ye Ren, Ponnuthurai N Suganthan, and Gehan Amaratunga. 2014. Ensemble deep learning for regression and time series forecasting. In *2014 IEEE symposium on computational intelligence in ensemble learning (CIEL)*. IEEE, 1–6.

[101] Aqsa Saeed Qureshi, Asifullah Khan, Aneela Zameer, and Anila Usman. 2017. Wind power prediction using deep neural network based meta regression and transfer learning. *Applied Soft Computing* 58 (2017), 742–755.

[102] Tobias Rachow, Sandy Berger, Michael Karl Boettger, Steffen Schulz, Salvador Guinjoan, Vikram K Yeragani, Andreas Voss, and Karl-Jürgen Bär. 2011. Nonlinear relationship between electrodermal activity and heart rate variability in patients with acute schizophrenia. *Psychophysiology* 48, 10 (2011), 1323–1332.

[103] Marko Radeta, Vanessa Cesario, Sónia Matos, and Valentina Nisi. 2017. Gaming versus storytelling: understanding children's interactive experiences in a museum setting. In *International Conference on Interactive Digital Storytelling*. Springer, 163–178.

[104] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–8.

[105] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Michael Thomas Knierim, and Alexander Maedche. 2018. Got Flow?: Using Machine Learning on Physiological Data to Classify Flow. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW612.

[106] Ma Mercedes T Rodrigo, Ryan SJ d Baker, Sidney D'Mello, Ma Celeste T Gonzalez, Maria CV Lagud, Sheryl AL Lim, Alexis F Macapanpan, Sheila AMS Pascua, Jerry Q Santillano, Jessica O Sugay, et al. 2008. Comparing learners' affect while using an intelligent tutoring system and a simulation problem solving game. In *International Conference on Intelligent Tutoring Systems*. Springer, 40–49.

[107] Steven M Ross and Gary R Morrison. 2004. Experimental research methods. *Handbook of research on educational communications and technology* 2 (2004), 1021–43.

[108] Isabelle Rouch, Pascal Wild, David Ansiau, and Jean-Claude Marquié. 2005. Shiftwork experience, age and cognitive performance. *Ergonomics* 48, 10 (2005), 1282–1293.

[109] L M Rudner. 2003. The classification accuracy of Measurement Decision Theory. In *Annual meeting of the National Council on Measurement in Education*. Chicago.

[110] Michael Russo, James McGhee, Edna Friedler, and Maria Thomas. 2005. Cognitive performance in operational environments. (2005).

[111] Michael B Russo, Melba C Stetz, and Maria L Thomas. 2005. Monitoring and predicting cognitive state and performance via physiological correlates of neuronal signals. *Aviation, space, and environmental medicine* 76, 7 (2005), C59–C63.

[112] Pasi Saari, Tuomas Eerola, and Olivier Lartillot. 2010. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 6 (2010), 1802–1812.

[113] Perry Sadorsky. 2006. Modeling and forecasting petroleum futures volatility. *Energy Economics* 28, 4 (2006), 467–488.

[114] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 32.

[115] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1997. Kernel principal component analysis. In *International conference on artificial neural networks*. Springer, 583–588.

[116] Michael C Schubert and David S Zee. 2010. Saccade and vestibular ocular motor adaptation. *Restorative neurology and neuroscience* 28, 1 (2010), 9–18.

[117] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. 2009. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine* 14, 2 (2009), 410–417.

[118] Kshitij Sharma, Zacharoula Papamitsiou, and Michail N Giannakos. 2019. Modelling Learners' Behaviour: A Novel Approach Using GARCH with Multimodal Data. In *European Conference on Technology Enhanced Learning*. Springer, 450–465.

[119] Yoshihiro Shimomura, Takumi Yoda, Koji Sugiura, Akinori Horiguchi, Koichi Iwanaga, and Tetsuo Katsuura. 2008. Use of frequency domain analysis of skin conductance for evaluation of mental workload. *Journal of physiological anthropology* 27, 4 (2008), 173–177.

[120] Shona E Simmons, Brian K Saxby, Francis P McGlone, and David A Jones. 2008. The effect of passive heating and head cooling on perception, cardiovascular function and cognitive performance in the heat. *European journal of applied physiology* 104, 2 (2008), 271–280.

[121] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[122] Erin T Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 4057–4066.

[123] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining Low and Mid-Level Gaze Features for Desktop Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–27.

[124] Richard T Stone and Chen-Shuang Wei. 2011. Exploring the linkage between facial expression and mental workload for arithmetic tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 55. SAGE Publications Sage CA: Los Angeles, CA, 616–619.

[125] Baohua Sun, Lin Yang, Patrick Dong, Wenhan Zhang, Jason Dong, and Charles Young. 2018. Ultra power-efficient cnn domain specific accelerator with 9.3 tops/watt for mobile and embedded applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1677–1685.

[126] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. 2015. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873* (2015).

[127] Emma VA Sylvester, Paul Bentzen, Ian R Bradbury, Marie Clément, Jon Pearce, John Horne, and Robert G Beiko. 2018. Applications of random forest feature selection for fine-scale genetic population assignment. *Evolutionary applications* 11, 2 (2018), 153–165.

[128] Hwan-Ching Tai and Dai-Ting Chung. 2012. Stradivari violins exhibit formant frequencies resembling vowels produced by females. *Savart Journal* 1, 2 (2012).

[129] Kaori Tamura, Tsuyoshi Okamoto, Misato Oi, Atsushi Shimada, Kohei Hatano, Masanori Yamada, Min Lu, and Shin'ichi Konomi. 2019. Pilot Study to Estimate Difficult Area in e-Learning Material by Physiological Measurements. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*. ACM, 35.

[130] Chek Tien Tan, Tuck Wah Leong, and Songjia Shen. 2014. Combining think-aloud and physiological data to understand video game experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 381–390.

[131] Timo Teräsvirta. 2009. An introduction to univariate GARCH models. In *Handbook of Financial time series*. Springer, 17–42.

[132] Julian F Thayer, Anita L Hansen, Evelyn Saus-Rose, and Bjorn Helge Johnsen. 2009. Heart rate variability, prefrontal neural function, and cognitive performance: the neurovisceral integration perspective on self-regulation, adaptation, and health. *Annals of Behavioral Medicine* 37, 2 (2009), 141–153.

[133] B Thon. 2015. Cognition and motor skill learning. *Annals of Physical and Rehabilitation Medicine* 58 (2015), e25.

[134] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[135] Simone Tognetti, Maurizio Garbarino, Andrea Tommaso Bonanno, Matteo Matteucci, and Andrea Bonarini. 2010. Enjoyment recognition from physiological data in a car racing game. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. ACM, 3–8.

[136] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global, 242–264.

[137] Hans PA Van Dongen and David F Dinges. 2000. Circadian rhythms in fatigue, alertness, and performance. *Principles and practice of sleep medicine* 20 (2000), 391–399.

[138] Kurt VanLehn. 1988. *Problem solving and cognitive skill acquisition.* Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY ....

[139] Athanasios Vourvopoulos, Evangelos Niforatos, and Michail Giannakos. 2019. EEGlass: An EEG-eyeware prototype for ubiquitous brain-computer interaction. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 647–652.

[140] Sandra Weintraub, Sureyya S Dikmen, Robert K Heaton, David S Tulsky, Philip D Zelazo, Patricia J Bauer, Noelle E Carlozzi, Jerry Slotkin, David Blitz, Kathleen Wallner-Allen, et al. 2013. Cognition assessment using the NIH Toolbox. *Neurology* 80, 11 Supplement 3 (2013), S54–S64.

[141] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 9.

[142] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[143] Christopher KI Williams and Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*. Vol. 2. MIT press Cambridge, MA.

[144] Daniel M Wolpert, Jörn Diedrichsen, and J Randall Flanagan. 2011. Principles of sensorimotor learning. *Nature Reviews Neuroscience* 12, 12 (2011), 739.

[145] David D Woods. 1985. Cognitive technologies: The design of joint human-machine cognitive systems. *AI magazine* 6, 4 (1985), 86–86.

[146] Grace C Wusk, Andrew F Abercromby, and Hampton C Gabler. 2019. Psychophysiological monitoring of aerospace crew state. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM*

International Symposium on Wearable Computers. ACM, 404–407.

[147] Xiang Xiao and Jingtao Wang. 2017. Undertanding and detecting divided attention in mobile mooc learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2411–2415.

[148] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 532–539.

[149] Robert M Yerkes and John D Dodson. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology* 18, 5 (1908), 459–482.

[150] Nick Yeung and Stephen Monsell. 2003. Switching between tasks of unequal familiarity: The role of stimulus-attribute and response-set selection. *Journal of Experimental Psychology: Human Perception and Performance* 29, 2 (2003), 455.

[151] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[152] HS Peter Yue and Rafi Rabipour. 1997. Methods and apparatus for noise conditioning in digital speech compression systems using linear predictive coding. US Patent 5,642,464.

[153] Roberto Zangróniz, Arturo Martínez-Rodrigo, José Pastor, María López, and Antonio Fernández-Caballero. 2017. Electrodermal activity sensor for classification of calm/distress condition. *Sensors* 17, 10 (2017), 2324.

[154] Xiao Zhang, Yongqiang Lyu, Xiaomin Luo, Jingyu Zhang, Chun Yu, Hao Yin, and Yuanchun Shi. 2018. Touch Sense: Touch Screen Based Mental Stress Sense. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 87.

7 APPENDIX A

7.1 Power spectral histogram:

The power spectrum of a time series describes the distribution of power into frequency components composing that signal. Once the frequency components are computed, they can be represented as a histogram (Power Spectral Histogram). We computed the mean, SD, skewness, kurtosis and median of the Power Spectral Histogram.

The average power of a signal is given by:

$$P = \lim_{T\to\infty} \frac{1}{T} \int_0^T |x(t)|^2 \, dt \tag{10}$$

To analyse the individual frequency component, we used the truncated Fourier transform and define the amplitude spectral density:

$$\hat{x}(\omega) = \frac{1}{\sqrt{T}} \int_0^T x(t) e^{-i\omega t} dt \tag{11}$$

from above the power density can be calculated using:

$$S_{xx}(\omega) = \lim_{T\to\infty} E\left[|\hat{x}(\omega)|^2\right] \tag{12}$$

where,

$$E\left[|\hat{x}(\omega)|^2\right] = \frac{1}{T} \int_0^T \int_0^T E[x^*(t)x(t')] e^{i\omega(t-t')} dt dt' \tag{13}$$

with x^* being the complex conjugate of x and t' provides the range granularity.

7.2 GARCH:

GARCH models are similar to AutoRegressive Moving Average (ARMA) models but they are applied to the variance of the data instead of being applied to the mean [4, 38, 69, 78, 113, 118]. GARCH processes $X(t)_{t\in\mathbb{Z}}$ take the general form

$$X_t = \sigma_t Z_t, t \in \mathbb{Z} \tag{14}$$

Where σ_t, the conditional deviance (so-called volatility in finance), is a function of the history up to time $t-1$ represented by H_{t-1} and $(Z_t)_{t\in\mathbb{Z}}$ a strict white noise process with mean zero and variance one. We assume that Z_t

is independent of H_{t-1}. Mathematically, σ_t is H_{t-1} measurable, where H_{t-1} is a filtration generated by $(X_s)_{s\leq t-1}$, and therefore

$$X_t|H_{t-1} = \sigma_t^2 \tag{15}$$

The series (X_t) follows a $GARCH(p, q)$ process if for all t

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{k=1}^{q} \eta_k \sigma_{t-j}^2, \alpha_j, \eta_k > 0 \tag{16}$$

The condition on the parameters, $\alpha_j = 1 \ldots p$ and, $\eta_k = 1 \ldots q$ for the GARCH equations to define a covariance stationary process with finite variance is that

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \eta_k < 1 \tag{17}$$

The rationale behind equation 16 is that, first, opposite to AutoRegressive Moving Average (ARMA) models, which are models for the conditional mean, the GARCH is a model for the conditional standard deviation. By "conditional" we mean "given the history up to time t", that is given H_{t-1}. Second, the model shows that more persistence is built into the variability. In other words, GARCH models the variance at time t in the time-series as the linear combination of the history of variances up to time $t-1$. For more details see [131]. The coefficients $\alpha_0 \ldots \alpha_p$ and $\eta_1 \ldots \eta_p$ can be estimated by maximizing a likelihood function. The most popular GARCH model is $GARCH(1, 1)$, that is, $p = q = 1$ in (3) meaning that the current action variability is explained by the latest action and the latest action number only (lag time of one).

7.3 LFSC:

LPC is susceptible to high peaks in the signal [7], hence we also compute the LSFC for the arousal data that improves upon this shortcoming of the LPC [67].

Following are the steps to compute the LFSC:

(1) Compute LPC. Let $\{a_i\}_{i=1}^m$ are the LPC coefficients.
(2) Compute the spectral Frequency using the following

$$\hat{Y}_m(\omega_k) = \frac{\hat{g}_m}{|\hat{A}_m(e^{jwk})|} \tag{18}$$

where, $\hat{g}_m$ is the prediction error of the m^{th} frame of the audio; and $\hat{A}_m$ is the Toeplitz normal equation [13] of order m.
(3) LSFC = $log|\hat{Y}_m(\omega_k)|$

7.4 LASSO:

To select the most important features we employ the Least Absolute Shrinkage and Selection Operator (LASSO) [134]. LASSO is an extension of Ordinary Least Square (OLS) regression techniques fit for the cases where the number of examples are less than the length of the feature vector [134]. To find the best fitting curve for a set of data points, OLS tries to minimize the Residual Sum of Squares (RSS) which is the difference between the actual values of the dependent variable (y) and the fitted values ($\hat{y}$). The formulation of the OLS is given as follows:

$$\hat{y} = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

The objective of the OLS regression is to minimize the difference between $\sum(\hat{y} - y)^2$ with the constraint that $\sum \beta_i^2 \leq s$. Where s is called the shrinkage factor. LASSO on the other hand performs similar optimization with the slight difference in the constraint, which is now $\sum abs(\beta_i) \leq s$. While using LASSO, some of the β_i will be zero. Choosing s is like choosing the number of predictors in a regression model. Cross-validation can be used to estimate the best suited value for s. Here, we use 10-fold cross validation to select the value of s. Our analysis seeks to identify how each of the extracted features from the different data-streams predicts participants' performance scores.

7.5 Random Forest:

Random forest is mostly used as a prediction algorithm, however, we will use it as a feature selection mechanism. random forests are ensembles of decision trees. The training algorithm for RF applies the general technique of bagging: repeatedly selects a random sample with replacement of the training set, fits trees to these samples, and uses these replicates as new testing sets. One of the key features of the random forest is that it can permute the given feature set and compute the feature importance for each feature in each dataset, by optimising one of the modelling parameters, e.g., root mean squared error, proportion of variance explained; or in the case of classifications, precision and/or recall. Using the individual feature importance from RFs, one can put a threshold either on the number of features or on the importance values of the features to select the required number of features.

7.6 Dimensionality Reduction

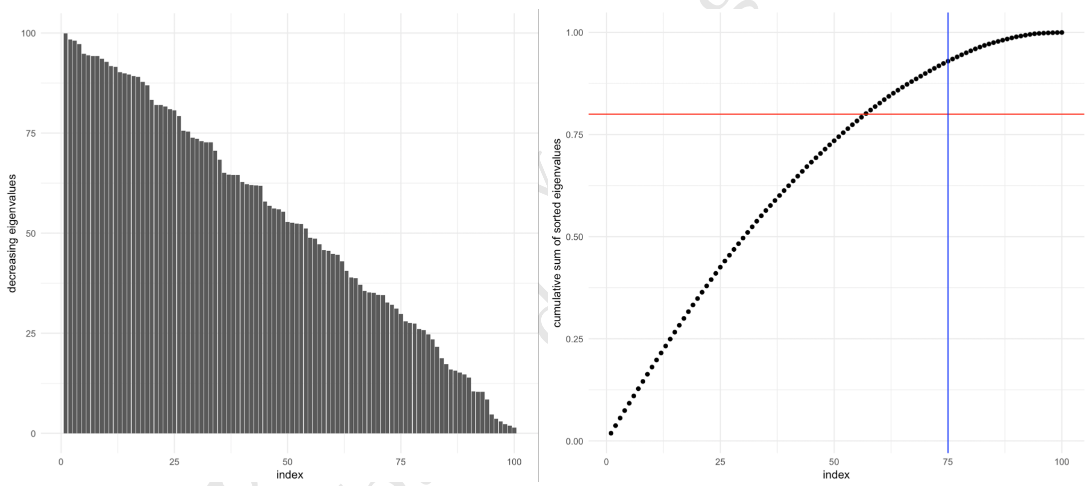

Fig. 15. Left: Simulated eigenvalues sorted in decreasing order. Right: Cumulative sum of the sorted eigenvalues; blue line is the threshold for the number of dimensions and the red line is the threshold for the percent of variance explained.

7.6.1 *Principal Component Analysis (PCA):*.

PCA identifies patterns that represent the data in a "better manner". The principal components could be seen as the new axes of the data maximizing the variance along those axes. This is achieved through the eigenvectors of the covariance matrix of the data. A common application of PCA is dimension reduction in a way that the information loss is minimised minimal loss of information. PCA projects the dataset (with d dimensions) onto a new subspace (k new dimensions where $k < d$). The main benefit of PCA is reduced computation time and also reduced error in the parameter estimation. PCA can be summarised in the following steps:

(1) compute the covariance matrix of the original data (X).
(2) compute the eigenvectors and eigenvalues of the original data.
(3) sort the eigenvalues in descending order (Figure 15 left panel).

(4) now, there are two different ways of reducing the number of dimensions in the original data. 1) pre-select the reduced number of dimensions and select the eigenvectors corresponding to the largest eigenvalues (see the blue line in the Figure 15 right panel). 2) put a threshold on the variance explained of the original data. This is equal to the proportion of the sum of the eigenvalues to the total sum of eigenvalues (see the red line in the figure 15 right panel).
(5) construct the projection matrix U using the k eigenvectors.
(6) project the data onto the new space using $Y = U^T \,.\, X$

7.6.2 Kernel PCA:. In the case where the data is not linearly separable, we would require a method to perform the dimensionality reduction using a way that considers the non-linear separation in the new space, since the linear dimensionality reduction will not yield good results. To perform the non-linear dimensionality reduction, we chose to use the kernel PCA, the basic working principle is the same as defined above, however we use a kernel function κ to compute the covariance matrix. The kernel is a function ϕ that transforms the data (d-dimensions) into a higher dimensional (p-dimensions) space, where the separation between the classes becomes linear again. Let us consider the sample X, the kernel function ϕ can be described as $X \to \phi(X)$. The individual data points in X would be projected to the higher dimensional space as follows (for details, see REF):

$$\kappa(x_i, x_j) = \phi(x_i)\phi(x_i)^T \tag{19}$$

For example, if X has two features

$$X = [x_i,\ x_j]^T \quad X \in \mathbb{R} \tag{20}$$

$$\downarrow \phi \tag{21}$$

$$X' = [x_1\ x_2\ x_1x_2\ x_1^2\ x_1^3x_2^2\ \ldots] \quad X \in \mathbb{R}^p (p >> d) \tag{22}$$

Next, to compute the covariance in kernel PCA, instead of using

$$Cov = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T \tag{23}$$

we use

$$Cov = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i)\phi(x_i)^T \tag{24}$$

7.7 SVM (Linear, polynomial, radial):

SVM maps an input X onto a multidimensional space using kernel functions (linear, radial or polynomial), and then any kind of regression can be used to model the input data in the new feature space (the kernel functions are described in the subsection "kernel PCA"). The quality of estimation is measured by the $\epsilon-$intensive loss function given by Chapelle and Vapnic (1992):

$$L_\epsilon(y, f(x,\omega)) = \begin{Bmatrix} 0 & if\ \ |y - f(x,\omega)| \le \epsilon \\ |y - f(x,\omega)| - \epsilon & otherwise \end{Bmatrix} \tag{25}$$

SVM regression performs regression in the high-dimensional space using $\epsilon-$intensive loss function, while minimising $\|\omega\|^2$. This can be achieved using non-negative slack variables to measure the deviation of training samples out of the $\epsilon-$intensive loss zone. The SVM tries to minimise $\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$ subjected to:

$$\left\{\begin{matrix} y_i - f(x,\omega) \le \epsilon + \xi_i^* \\ f(x,\omega) - y_i \le \epsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0,\ i = 1..N(slackvariables) \end{matrix}\right\} \tag{26}$$

this can be transformed to

$$f(x) = \sum_{i=1}^{N_{sv}}(\alpha_i - \alpha_i^*)\kappa(x_i, x) \quad subject\ to\ \ 0 \le \alpha_i, \alpha_i^* \le C \tag{27}$$

Where N_{sv} is the number of support vectors and κ is the kernel function.

7.8 Model tree M5:

These are based on decision trees, which let us split the data into separate smaller datasets or "islands" using different feature sub-spaces. The main purpose of such splits is to minimise the overall weighted loss on the data. What is commonly used in decision tree classification is the mean-regression with L2 loss for decision tree regression. Model Trees extend the decision trees by allowing us to build decision trees out of any model of our choice.

7.9 Gaussian process model (Linear, polynomial, radial):

This model is like SVM, the only difference being the fact that the mapping from the original space to a multidimensional space is governed by Gaussian latent variables that are parametrized using different kernel functions (Rasmussen and Williams, 2016).

$$P(Y|X,\theta) = \prod_{i=1}^{D} \frac{1}{(2\pi)^{\frac{1}{2}}\,|\kappa|^{\frac{1}{2}}} e^{-y_i^T\kappa^{-1}y_i} \tag{28}$$

Where θ is the set of hyperparameters, κ is the kernel function. D is the dimensions of the original data X and Y is the target variable. In this study, we set the kernel functions to take linear, polynomial and radial forms.